Flinders UNIVERSITY
ADELAIDE • SOUTH AUSTRALIA

Critically thinking about critical thinking in science education:

Interrogating the perceptions and actions of Australian senior secondary and tertiary educators

by

Amy Butler

Thesis
Submitted to Flinders University
for the degree of

Doctor of Philosophy
College of Science and Engineering
March 2020

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

LIST OF ABBREVIATIONS

ACARA	Australian Curriculum, Assessment and Reporting Authority
ACER	Australian Council for Educational Research
ACT Framework	Adaptive Critical Thinking Framework
AHELO	Assessment of Higher Education Learning Outcomes
ALE	active learning environment
AUQA	Australian Universities Quality Agency
B&S	biology and society
CAT	Critical-thinking Assessment Test
CC	course coordinator
CLA	Collegiate Learning Assessment
CLA+	Collegiate Learning Assessment Plus
COPUS	Classroom Observation Protocol for Undergraduate STEM
CT	critical thinking
EQUIP	Electronic Quality of Inquiry Protocol
EWCTET	Ennis Weir Critical Thinking Essay Test
GPA	Grade Point Average
HS	high school
MCQ	multiple choice question
PCoA	Principal Coordinates Analysis

OECD	Organisation for Economic Co-operation and Development
QSRLS	Queensland School Reform Longitudinal Study
SALG	Student Assessment of their Learning Gains
STEM	science, technology, engineering and math
TA	teaching assistant
TEQSA	Tertiary Education Quality and Standards Agency
TER	tertiary (in reference to university/college level studies)
US	United States of America

GLOSSARY

Active learning:

This term is applied in a number of ways in the literature, in the context of this thesis it is used to imply learning opportunities that require a student-centred and directed action (such as asking or answering a question, working in a group, participating in class discussion etc.). See Chapter 4 for an explanation of how active learning opportunities were captured through the classroom observation protocol I deployed.

Adaptive Critical Thinking (ACT) Framework definition:

"Critical thinking is a purposeful inquiry process that involves the deployment of thinking skills and context-appropriate criteria to make reasoned, reflective judgments and transform information into useable knowledge to resolve points of uncertainty" (Butler, 2019 see Chapter 2).

American Philosophical Association definition of critical thinking:

"We understand critical thinking [CT] to be purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based. CT is essential as a tool of inquiry" (Facione, 1990a, p.2).

Discipline-Based Education Research (DBER):

An interdisciplinary research approach often associated with exploring teaching and learning in science, engineering and math. The research employs social science methods to explore classroom practice and outcomes (Singer, Nielsen, & Schweingruber, 2012). It aims to generate understanding that is not limited to a classroom by providing transferrable "insights into educational processes and their effects" (Dolan et al., 2017, p.2).

Embedded approach:

This approach refers to embedding critical thinking learning opportunities within the classroom. According to Ennis this teaching approach could involve the infusion or immersion of critical thinking in a classroom (Ennis, 1989 - see also definitions below); however, in Australia (and the context of my thesis), embedded approaches usually mean that critical thinking related language is not explicitly used (which Ennis would term an immersion approach).

Explicit approach:

This approach refers to the explicit use of critical thinking related terminology and language concerning skills and disposition when incorporating critical thinking in a classroom (Abrami et al., 2015). This approach is different from the general approach because the emphasis is on the language around critical thinking, not the content (or lack thereof) used to demonstrate or scaffold critical thinking.

Infusion approach:

This approach involves explicitly teaching critical thinking general principles in a discipline-specific environment while encouraging students to deploy these attributes on discipline-specific content (Ennis, 1989). Critical thinking is a specific and explicit objective of the course (Abrami et al., 2015).

Immersion approach:

This approach involves training students in discipline-specific/context-specific critical thinking (in other words to deploy critical thinking skills and dispositions on disciplinary content) without making the general critical thinking principals explicit (Ennis, 1989). Unlike the infusion approach, critical thinking is not an explicit objective of the course.

General approach:

This approach involves teaching critical thinking explicitly and separately from content (in other words, without being anchored in any one discipline area or subject matter). This approach has an emphasis on general and transferable nature of the logic of critical thinking (Ennis, 1989).

Mixed approach

This technique is a combination of the general approach combined with discipline-specific content and other examples. Ennis describes this approach as a combination of the infusion and immersion approach (Ennis, 1989). This approach typically leads to the strongest CT development outcomes (Abrami et al., 2008, 2015).

Principal Coordinates Analysis (PCoA):

Principal Coordinates Analysis is a visualisation method for exploring differences in multivariate data which has first been transformed into values in proximity matrices (Clarke and Warwick, 2001).

Scholarship of Teaching and Learning (SoTL):

A common higher education-based reflective practice approach associated with exploring and improving teaching and learning. This process aims to explore, generate and employ localised evidence-based insights for improving classroom outcomes (Shulman, 2000; Shipley, McConnell, McNeal, Petcovic and John, 2017).

Threshold Learning Outcomes:

Nationally-developed approach for expectations around the minimum level of achievement required for a bachelor degree in Australia, where each discipline has its own set of threshold standards (Jones, Yates and Kelder, 2011). Outcomes were developed in conjunction with advisory committees and feedback from relevant disciplinary communities including academics, employers and professional societies across Australia.

ABSTRACT

Critical thinking (CT) is an essential skill for the workplace and education (Oliver and Jorre de St Jorre, 2018; Sellars, 2018). Yet globally, educators feel ill-prepared and ill-equipped to foster CT in students (Lauer, 2005; Choy and Cheah, 2009; Black, 2009; Phelan, 2012; Aliakbari and Sadeghdaghighi, 2013; Reynolds, 2016; Carbone et al., 2019). Despite multiple findings confirming that mixed teaching approaches generate the strongest CT development outcomes (Tiruneh, Verburgh, and Elen, 2014; Abrami et al., 2015), instructional methods continue to emphasise embedding CT into curricula design (Puig, Blanco-Anaya, Bargiel and Crujeiras-Pérez, 2019). This is especially true in Australia, where an embedded approach is advised through policy documents. There are ongoing calls for classroom-based research into 21st-century skills such as CT, with a demand for information that reveals best-practice approaches (Arum, Roksa and Cook, 2016; Doeke and Maire, 2019). Concurrently, CT assessments are criticised for failing to help educators understand how classroom dynamics impact students' CT development (Benjamin, 2012; Rear, 2019). Consequently, the purpose of this doctoral research is to empower educators with knowledge and tools to increase CT development in their classrooms.

This transdisciplinary thesis contributes empirical, methodological and theoretical knowledge about CT, with a focus on its role in science education. In respect to the empirical contribution, this research adds to understanding of CT development in Australia through an exposition of educator ideas, tertiary science teaching activities and CT development outcomes. It includes the first cross-disciplinary, multi-institutional survey relating to educator perspectives of CT in Australia. This doctoral research makes two methodological contributions by applying scientific thinking to education research. Firstly, it demonstrates how a biologically-based statistical approach can generate deeper understanding of survey data. Secondly, it demonstrates how using a multi-instrument analysis approach creates deeper and actionable understanding of CT development. The main theoretical contribution of this doctoral research – the Adaptive Critical Thinking Framework – is an original synthesis and reconfiguration of the components of CT. Descriptions of the framework elements characterise the essential elements of CT. Elaborations of this framework, including the accompanying definition, summarise the process of CT and highlight the importance of context when applying the elements. This framework is applied throughout the thesis to exhibit its capacity to enhance understanding about CT.

Collectively, this research supplies new insights into CT development in Australia. It demonstrates the capacity of a new framework and a multi-tool assessment approach to enhance understanding of opportunities for CT development in classrooms. Contemporary science graduates require more than just standard scientific skills and content knowledge to succeed (Taber, 2016; Pearl, Rayner, Larson, and Orlando, 2019). Employers require graduates to be reflexive thinkers and problem-solvers (Bezanilla, Fernández-Nogueira, Poblete and Galindo-Domínguez, 2019). Yet the way to achieve this outcome requires strategic changes to teaching practice. Educators need clear policy direction and training to support their decisions and efforts to enhance CT development in their classrooms. There is a need to overcome the failure to measure what matters (Shively, Stith, and Rubenstein, 2018). It is time to think more critically about critical thinking.

DECLARATION

I certify that this thesis: does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and to the best of my knowledge and belief, does not contain any material previously published or written by another person except where due reference is made in the text. I confirm that this thesis has been written and edited without the services of a professional editor. I have completed all writing and editing of this thesis.

Signed Amy Butler

Date 17/11/2019

ACKNOWLEDGEMENTS

I am blessed to have encountered many wonderful people through this thesis journey. Words hardly seem enough to thank them for their time, advice and friendship. I aspire to emulate the best qualities from each of you and am grateful you saw potential in me. Please know that you have helped shape me into a more thoughtful, generous and confident soul.

To my supervisors - thank you for investing in me. Martin, thank you for inspiring me to try and change the world. Your endless questions always challenged me to think until it hurt. I can still hear your voice prompting me to consider the purpose of all that I do and write. Karen, thank you for seeing potential in me as an undergrad and for providing opportunities to develop as an academic and researcher. Jeanne, you have been a constant mentor and helped me through so many moments in this long and hard journey. I would not have made it this far without you and your feedback and will value your friendship forever. Tara, I am incredibly appreciative for your advice and your belief in me. You lift the spirits of those around you and help us aspire to be more. Thank you for understanding what I was trying to achieve with my research and for helping me crystallise the path I wanted to travel.

To the amazing OGR staff - thank you for being fabulous. The changes you implement make our journey so much easier. Ashleigh, thank you for helping me navigate moments where it would have just been easier to walk away. Lisa, thank you for creating openings for me in Tara's hectic schedule.

To my Science21 family - each of you have helped me understand a little more about the challenges we face in changing the culture of education. It was always inspiring to hear about the work you do, and to consider the different ways that content can be presented to enable deeper learning. While we all may be walking different paths now, I will be ever grateful for your insights and support.

To my favourite academics: JQ, Narelle, Ingo, Claire and Chris. Thank you for your love of teaching and learning. Narelle, Ingo, Claire and Chris, I appreciate your willingness to allow me to evaluate your courses, and for the time you invested in scoring the CAT tests. JQ, from that first open day we met - you got me. From undergrad you have been my champion. Thank you for your sage advice and for always being so generous with your time.

To the Wos-Oxley family - Mel, you appeared in my life at just the right time. Thank you for inspiring me to love the story of numbers. I will forever appreciate all those hours you spent teaching me statistics. I am grateful for the brain breaks with Freddo and Ida. Andrew thank you for releasing Mel from parental duties through all those nights we stayed up late crunching numbers.

Thanks to the Australian Government for a Research Training Program Scholarship to support me during this journey.

To my dear friends: Louise, Linda, Mel, Sam, Deb, Jen, Lauren, Rach, Nads, Cat, Amy, Yas and Claudz. Thank you for sticking by me through my sporadic availability and for the constant reminders that there is life outside of thesis writing. You have all kept me sane through this journey.

To my family - thank you for listening to my ideas and protests. Mum – thank you for never complaining about the number of drafts I asked you to read, and for all of the feedback you provided. Dad - thank you for being a soundboard and for storing those boxes a lot longer than we thought! Matt and Kalani – thank you for always checking in with me, it means more than you could ever know!

To my darling Callum - where would I be without you? I am ever grateful of the fact that you always allow me to pursue my dreams. You keep me grounded and feed me before I get hangry. Thank you for all the emotional and financial support you have provided throughout this journey. But most of all, thank you for your patience.

Introduction:

From Classroom to Concept

The opening up of new points of view and new methods are inherent in the progress of knowledge.

Dewey, 1929, p.31

Recognising the importance of critical thinking (CT) is not new (Ennis, 1962; Brabeck, 1983; Giancarlo and Facione, 2001; Grieco, 2016; van der Zanden, Denessen, Cillessen and Meijer, 2018a), with early conceptions about CT stemming from Greek philosophy (Paul, Elder and Bartell, 1997a). Peirce (1877) also contributed understanding about inquiry, belief and the role of doubt in thinking. He reflected on the work of Lord Roger Bacon and scientists such as Kepler, Lavoisier and Darwin, and noted that the memorable works of science were a "lesson in logic" and "reasoning" (Peirce, 1877, p.2). However, it is Dewey (1910), Glaser (1941) and Ennis (1962) who are considered founders of the modern understanding of CT (Fisher, 2011).

Many influential contributors have transformed the way CT is incorporated into education programs (McPeck, 1981, 1990; Glaser, 1983, 1984; Norris, 1985, 1989; Sternberg, 1986; Facione, 1990a, 1990b; Paul, 1995, 2005, 2011; Nosich, 1996; Paul and Elder, 2006a; Halpern, 1998, 1999, 2003). Of importance to this doctoral research is the stance put forward by the American Philosophical Association (Facione, 1990a, 1990b). Generated by a panel of experts seeking to clarify the role of CT in education, this view considers,

> critical thinking [CT] to be purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based. CT is essential as a tool of inquiry. As such, CT is a liberating force in education and a powerful resource in one's personal and civic life. While not synonymous with good thinking, CT is a pervasive and self-rectifying human phenomenon.
>
> (Facione, 1990a, p.2)

Yet despite Facione's efforts, and the numerous other contributions from theorists from philosophy, psychology and education, CT remains a problematic concept. Examination of the literature highlights ongoing calls for a consensus on CT (see Poirier and Hocker, 1993; Capossela, 1998; Sanders and Moulenbelt, 2011; Lai, 2011a; Moore, 2013; Dunne, 2015). As a result, definitions have continued to be generated (see Halpern, 2003; Paul and Elder, 2006a; Van Gyn and Ford, 2006; Paul, 2011, 2012; Ennis, 2011a; Dwyer, Hogan and Stewart, 2014; Thomas and Lok, 2015).

In Australian education, expectations about CT have been criticised as being too vague for students, particularly international ones (Vandermensbrugghe, 2004). Policymakers in the Australian education system have sought to acknowledge the role of critical and creative thinking by incorporating aspects of these thinking skills in key policy documents, such as the Australian Curriculum and in the tertiary education Threshold Learning Outcomes (Jones et al., 2011; ACARA, 2013). In these documents, desirable skills sets and dispositions are listed in an attempt to provide a functioning definition of what CT looks like. However, this is problematic because focussing on the skills decontextualises them. It makes them seem like the essential component of CT, when really it is the ability to think critically and the development of good mental habits which underpins success in these cognitive skill areas (Costa and Kallick, 2000). Further, various scholarly disciplines place emphasis on different sets of thinking processes (Gordon, 2000). It is not so much the possession of any one skill which is important when it comes to CT; but the employment of a range of appropriate cognitive thinking skills to end up with a quality response for that context (Bailin, 2002). Therefore, views about CT that focus on individual parts of the CT process in isolation are too reductive for their application to amount to a truly critical thought product.

The volume of literature on CT development since the 1990s has increased from a handful of papers to thousands. This ever-growing range of perspectives on CT presents a challenge for educators. There is so much information to process, and not all perspectives demonstrate a sound evidential base. Commentators across various fields note that much of the research tends to be qualitative in nature, using anecdotal or self-report data to provide evidence of student gains in CT (Staib, 2003; Leming, 2016). While some disciplines will accept these kinds of results as a validation of the approaches effect on CT development, others, such as science, generally require a more quantified approach. However, the value of studies where quantitative measures are used also varies. Most current assessment approaches isolate each skill to reveal strengths and weaknesses (Ennis, 2009), but given the intertwining nature of CT skills, treating skills in isolation is likely to be to the detriment of understanding the quality of thought (Benjamin et al., 2013). Further, the tools used to qualify the approaches are often too general and disconnected from the context to provide an educator with a specific understanding of CT development in their classroom (Benjamin, 2012). What remains is a failure to achieve a consensus stance about the concept of CT and failure to clarify CT development in contextually relevant ways for educators who want to work from an evidence-base perceptive on CT.

A practical solution is to select and deploy an approach and continue with a trial and error process until a successful outcome is found for their classroom. However, findings from previous perception studies indicate it is unlikely educators have the time or training to determine how to overcome weaknesses in existing perspectives when making judgements about the best methods to use. For example, Shell (2001), Black (2009) and Aliakbari and Sadeghdaghighi (2012) all found that time constraints and a lack of training and support for educators were the major barriers impeding the incorporation of CT in classrooms. Even Hatcher (2013) recognised that the detailed explorations and evaluations undertaken throughout his program development would not have been possible without 14 years of funding (cumulatively worth over a million dollars). This funding allowed Hatcher to reduce teaching loads, educate teachers through conferences and training, and bring in experts to help with program design. Few educators have access to such resources. Therefore, to overcome barriers around training and time constraints, tools for navigating the literature and guiding educator judgements are needed. These tools not only equip educators with the resources they need to teach well, but also help them understand the what, the why and the how so they can align their existing ideas to evidence-based approaches. However, an understanding of existing knowledge and practice is needed to improve the relevance of these tools.

Perspectives from educators, from several countries and disciplines, have already been examined (see Paul et al., 1997a, 1997b; Choy and Cheah, 2009; Kowalczyk, Hackworth and Case-Smith, 2012; Stedman and Adams, 2012; Rodzalan and Saat, 2015). However, there is a general gap in the understanding of perspectives at a K-12 level, especially in science, because aside from Black (2009) and Alwadai (2014), previous CT perception studies have focused on perspectives from higher education. There is also a lack of knowledge about Australian educators, as only one study, from one faculty of history, philosophy and cultural studies educators has been undertaken (Moore, 2013). There is also a lack of classroom-based data on CT development in science classrooms, as momentum for Australia's STEM movement is fairly recent (Murphy, MacDonald, Danaia and Wang, 2019). My research addresses these gaps in knowledge by examining the views Australian high school and university science educators hold regarding the definition, development and assessment of CT. However, the reliability of perception and self-report data is limited and does not always accurately represent classroom outcomes (Paul, 2005; Beck and Blumer, 2016; Leming, 2016). For example, Paul (2005) expressed that some educators think "they understand critical thinking sufficiently and are already successfully teaching it" (p.27), even when this is not the case. However, students can be equally as poor at assessing their learning (Porter, 2013; Halpern, 2014). Therefore, to build on my educator perception survey findings, this research also investigates science educator actions using a case study approach.

Leming (2016) suggests that educators draw from their prior education experiences to shape the pedagogical choices they make. As a scientist, I have learnt to trust in evidence-based knowledge, and have been trained to observe, hypothesise, quantify and draw conclusions from new data. These habits have carried through to the way I function as a science educator. I want to inspire and help students understand the world around them using teaching techniques that have a quantified evidence-base. However, when it comes to CT I have not been able to find evidence that satisfies the scientist in me, nor have I encountered many students who implicitly possess the processing skills needed to navigate the plethora of information that we now have at our fingertips.

As I reflect on my education journey, I realise that I have been privileged. I have been given permission to remain inquisitive - from participating in gifted learning programs where deep thinking was endorsed, to simply being exposed to educators who allowed and encouraged me to ask why. I am a better thinker because of it, and even as an adult my curiosity and love of learning remains. But this is not the typical pathway for Australian students. At least not those I have observed through my various experiences working with children – which range from nurturing babies through to teens in various childcare environments, to working in schools with the CSIRO and teaching university students. These experiences have positioned me to notice that our experiences with knowledge shape our capacity to think. We encourage and allow a young child to ask why as they start to develop their understanding of the world, but at some point in their education journey, it becomes unacceptable to question the knowledge that an adult tells them. There remains too much emphasis on retaining knowledge rather than understanding how to navigate and apply it. As technological advances continue to shift social and economic structures from an industrial to a knowledge-based economy, education objectives need to change (Thompson, 1967; Etzkowitz, Schuler and Gulbrandsen, 2000; Lautensach, 2004). Continuing to impart 'content knowledge' at the expense of developing thinking skills will not help meet the needs of the future workforce (Ironside, 2004) nor help solve issues relating to climate change and resource depletion (Lautensach, 2017). In this post-truth world (Gross, 2017), we need information-literate thinkers. However, to resolve these issues and challenges, increased understanding of CT development in Australia is needed. To increase information literacy and thinking skills in graduating students, it is necessary to engage them and train them to appropriately challenge and verify knowledge (Hirsh, 1997; Rieh and Hilligoss, 2008; Case and Given, 2016). These skills are at the core of critical thinking.

Research scope, aims and questions

Examination of CT literature revealed extensive information for educators to interpret, evaluate and understand. However, theoretical constructs are not always evidence-based or well-suited to classrooms. In addition, educators across the globe generally do not have enough time or training to be able to think critically about CT for themselves. There is a paucity of literature on Australian educators, particularly within the sciences, despite the development of CT (or at least CT skills) being an intended learning outcome from the Australian National Curriculum and also from the Threshold Learning Outcomes that help guide university graduate standards (Jones et al., 2011; ACARA, 2013; Oliver and Jorre de St Jorre, 2018). There is also no existing literature outlining the trends in Australian science educators thinking, and their actions when it comes to CT development. Since so many of their international peers are feeling ill-prepared and unsupported in this area, it is important to gain insights into Australian educators' practice so we can understand how to support them. However, if the variety of perspectives on CT in the existing literature is anything to go by, this is not a simple problem to address.

Transdisciplinary research is a problem-solving approach that starts with real-world issues (Wickson, Carew and Russell, 2006) and crosses discipline boundaries to help unify and generate knowledge in a way that cannot be achieved within one disciplinary space (Mahan, 1970; Hadorn et al., 2008; Bernstein, 2015). Transdisciplinary inquiry needs CT (Wagner, Baum and Newbill, 2014). This movement of knowledge across disciplines requires intense thinking, reflection and judgement - all of which are essential mental processes in CT. However CT also requires transdisciplinary research. For too long CT has encountered translational problems about its construct, process and development. There are calls for consensus, even though numerous theorists have supplied collective perspectives and frameworks. There are demands for practice-based evidence so that the theories being generated about CT align to real-world experiences. Yet these calls occur concurrently with requests for examples of evidence-based practice from educators who want to be able to 'get on with' developing students' CT abilities using verified approaches. This response is due to policymakers and education institutions demanding CT as a learning outcome even though the educators at the student/learning interface are not necessarily trained (nor supported) to facilitate CT development. Each stakeholder requires more than current understanding about CT can supply. To help solve these problems, a new approach is needed.

My transdisciplinary research aligns the areas of science, CT and curriculum development to enhance understanding about CT development in science in Australia (see Figure 1). It incorporates some of the body of work on CT to seek out commonalities across the field and then translates them into a framework that can be adapted for widespread use. It incorporates existing understanding about curriculum development focusing on student learning and CT to reconfigure methods for CT development and assessment. However, it also consults important stakeholders (Australian science educators) to increase the relevance of the research for the Australian context. Thinking about CT in the context of curriculum development is not new. In fact, it is what most of the literature on CT development is about. However, what is new to this transdisciplinary nexus is science. Science is largely responsible for the mass knowledge generation in the modern age. Yet the knowledge science produces is often fragmented - a by-product of the discipline-based approach science tends to employ (Hoffmann-Riem et al., 2008). Scientific approaches are long thought to contribute skills akin to those associated with CT (Dewey, 1910; Byrne and Johnstone, 1987; Klahr, 2000). But most importantly, science places an emphasis on evidence-based understanding, which resonates with calls for more evidence on CT. Given the contributions of science to the information literacy demands stemming from a knowledge-based economy, it seems appropriate that scientific thinking and methods are part of the solution to resolve it.

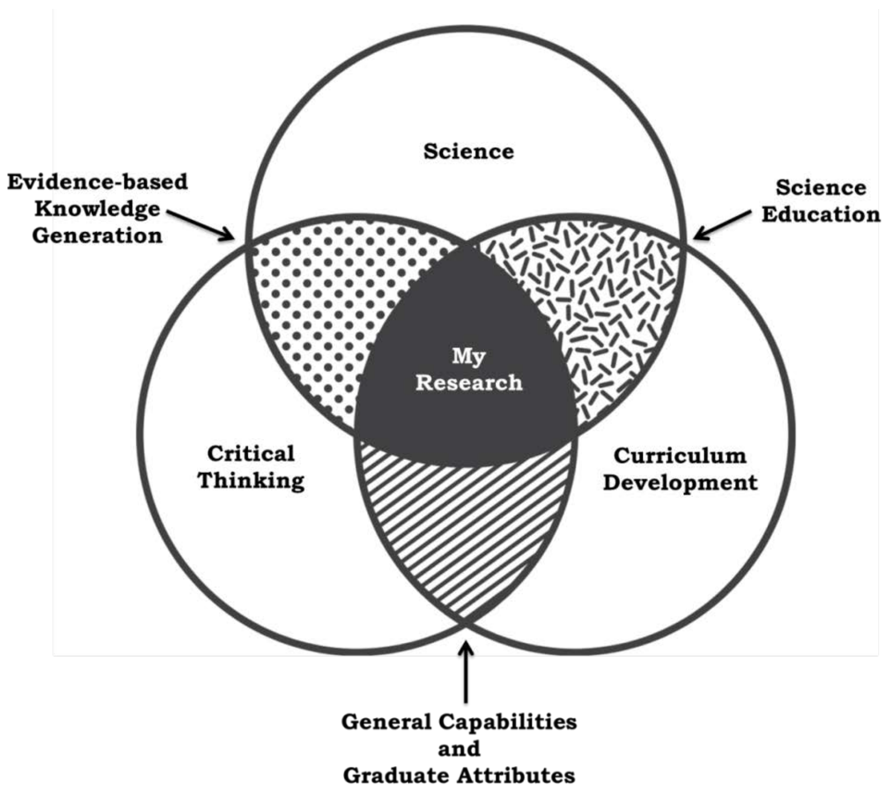


Figure 1 My transdisciplinary research frame. Figure created for Butler (2019).

This transdisciplinary doctoral research has three aims:

1. To synthesise the theoretical construct of CT, to clarify the literature and generate an adaptable framework for CT and increase the accessibility of ideas on CT.
2. To investigate Australian educator perceptions about CT and its development and assessment, and determine whether their understanding is consistent with Australian policy and/or the literature on CT.
3. To explore how Australian tertiary science educators at Flinders University are developing CT and determine the effect of their approaches using a case-study method that could also serve as a general model for CT assessment.

The first aim of this research is to the address the usability of the literature on CT, incorporating the iterative and emerging definitions. This was achieved by generating an adaptable CT framework and a supporting definition to ensure the effective deployment of the framework. This CT framework also became a way to further analyse the data gathered for aim 2 and 3. Concerning aim 2, it was used in terms of determining which aspects of the framework are present in educator perspectives. Concerning aim 3, it was used to identify areas of weakness and highlight opportunities for improvement within student CT development

To address the lack of an Australian understanding on CT (aim 2) and explore the success of CT development in science (aim 3), the research investigated the perceptions and actions of Australian science educators regarding CT. Aim 2 was achieved by exploring the views that Australian high school and university science educators held about the definition, development and assessment of CT. Then to further investigate the development of CT, and to contribute insights to assist educators with making development-based and assessment-based decisions in their classrooms (aim 3), the second part of the research measures, clarifies and compares the effects of different CT development approaches in physical and life sciences.

The main research questions concerning aims 2 and 3 are:

1) **What ideas are represented in Australian educators' perceptions of critical thinking and what approaches do educators report incorporating in their classrooms to develop and assess CT?**

    a. What do Australian science educators know about the nature of CT, and is their understanding different depending on their discipline or education setting?

b. What approaches do Australian educators report incorporating in their classrooms to develop CT, and are these approaches different depending on discipline or education setting?
c. What methods and tools would Australian educators use to assess CT, and are these different depending on discipline or education setting?
d. Do factors such as teaching experience, state (as a proxy for identifying state-based policies), or gender, help explain differences in educator perceptions about CT?

2) What teaching approaches are currently employed by Australian science educators in order to develop critical thinking abilities, and which are most effective?

To explore CT development in science, this research deploys a combination of science education research practices, including Scholarship of Teaching and Learning (SoTL) and Discipline-Based Education Research (DBER). SOTL is a reflective teaching process that engages with evidence-informed approaches (both literature and data from one's own classroom) with an emphasis on improving classroom learning outcomes (Shulman, 2000). Whereas DBER is often described as a form of interdisciplinary research, using methods grounded in social science to help address "discipline-specific problems" that arise when trying to teach evolving scientific knowledge using evidence-informed approaches (Singer et al., 2012, p.202). The key difference between these approaches is the purpose of the research. The SoTL process tends to be an individualised and localised evidence-based learning approach, seeking to inform practice at a classroom level (Shipley et al., 2017). Whereas the goals of DBER are broader with the research seeking to generate understanding that extends beyond a "single classroom" or "program" to provide "insights into educational processes and their effects" (Dolan et al., 2017, p.2).

Since transdisciplinary research exposes and transcends traditional disciplinary boundaries, an important part of the work is identifying the disciplinary interactions and the overall research frame (Mahan, 1970; Bernstein, 2015). Between my framework, the perception survey and the case studies, this transdisciplinary research is intended as a launching point for understanding CT development in senior secondary and tertiary science classrooms in Australia. However, there are many facets of science, curriculum development and CT that were outside the scope of the research (see Figure 1). For example, the research focuses on CT and curriculum development through the lenses of science and Australia. Whilst perspectives were initially gathered from

Australian science and history educators, most of the classroom-based research component focussed only on those modalities which were appropriate to science. Another boundary for the curriculum development area is how my research tackles CT assessment. I sought to collate and apply CT assessment tools in a unique way to increase the usability of classroom data for making change, but this research is not intended to address the verification or validation of new or existing measures of CT. Rather, it is intended to show a possible pathway through to creating informative and usable evidence-based research on CT. The boundary relating to CT and my research is less overt. In seeking to synthesise the literature on CT to formulate boundaries relevant to curriculum development in science, larger problems were revealed. The core of which, for me, was that educators were faced with a plethora of varying perspectives on CT and were ill-equipped to deal with it. Perhaps ironically, this resulted in the generation of another perspective. However, my framework is not intended to replace existing understanding about CT. Rather it is meant to enhance understanding by making the literature more accessible. As such, the boundary between my research and CT is a little more fluid. It set out to take existing understanding and apply it, but because of the intense interactions between using my training as a scientist and an educator to think about CT, it resulted in a new way of knowing about CT.

Explanation of important terms[1]

For the purpose of this research, the term 'science educator' encompasses physics, chemistry and biology teachers in both tertiary education and senior secondary education. These educators teach courses that require discipline expertise, unlike educators from lower levels of schooling, where approaches are much more general. This choice meant investigations were informed by the discipline-specific perspectives that these educators potentially held and allowed for exploration of policy-related perspectives. In addition to investigating science educators, non-science educators from history were included in initial investigations. These history educators functioned as an outgroup would in phylogenetics, helping to determine if there were unique perspectives in the science educator group, or if the perspectives were common to both sciences and humanities. History was selected purely on the basis that there were comparable subject offerings at a senior secondary and tertiary level, in addition to the fact that CT is currently outlined as a learning outcome in both the Australian Curriculum and tertiary Academic Standards.

[1] Other important terms and definitions can be found in the glossary.

This meant that both educator groups would have some discipline expertise, as well as policies that could potentially guide their practice. Thus, in the context of this study, ‘Australian educator’ refers to a senior secondary or tertiary practitioner from science (physics, chemistry or biology) or from a non-science subject (history). Further, senior secondary educators are referred to as high school educators. An additional education-related term to note is that the word ‘course’ is used throughout this research. For the purposes of this study the term ‘course’ should be understood as a single semester of teaching (as per the USA). In Australia, this would normally be referred to as a 'topic,' with the word course being used in reference to the entire degree program.

This research applies a scientific analysis approach to data that would often be handled qualitatively or by deploying basic statistics (including descriptive and univariate methods). As such, it is worth considering the difference between descriptive, univariate and multivariate approaches. Descriptive statistics use measures that consider one variable at a time. In contrast, univariate and multivariate analyses explore the relationship between two or more variables (Anderson, 2003; Lehamn, O’Rourke and Stepanski, 2005). The key point of difference between univariate and multivariate analyses is that univariate statistics generally involve the exploration of one dependent variable at a time, whereas multivariate statistics involves concurrently studying two or more dependent variables. However, by function, univariate is used to encompass all comparisons which can be handled by an analysis of variance (ANOVA). This reserves the term multivariate statistics for those analyses that have multiple dependent variables (and/or factors) that cannot be handled by an ANOVA and/or be otherwise analysed through parametric approaches.

Chapter structure and outlines

The following section overviews the contents of the thesis chapters. This body of research consists of two theoretical chapters (Chapters 1 and 2) which explore the construct of CT, and four experimental chapters (Chapters 3-6). These experimental chapters contain the study methods; some additional literature explanations related to the assessment and analysis choices; and the results and findings concerning educator perceptions and actions in relation to CT development.

There is a key structural intervention in this doctoral thesis. Because of the transdisciplinary imperative, it became important to ensure the connection between the chapters and reflect on their relationship. My research explores diverse but important ideas about CT and its development in Australia, so these reflections function like a primer in DNA synthesis, providing the starting point for the next stage of the synthesis. They also reveal my transdisciplinary journey from thinking like a scientist to thinking like a science educator informing curriculum development. Therefore, these sections - labelled **primers** - intervene and interrupt the argument, opening spaces and offering intellectual challenges.

Chapter 1: Literature Review – This chapter contains a review of the existing definitions and frameworks for CT. It explores why, despite hundreds of years of theorising, a consensus on CT has still not been achieved. It considers the role of context and criteria in shaping the CT process. This chapter also describes the challenges faced by educators when attempting to use the existing literature and incorporate CT in their classrooms.

Chapter 2: Adaptive Critical Thinking Framework – This chapter presents a synthesis of the literature led to the construction of a CT framework. The framework then became a tool for not only explaining the construct of CT but also a way to investigate it. This chapter presents the framework and supporting definition I developed to help increase the usability of the literature and bring a shared understanding of CT. It examines the process of CT and the core aspects needed to differentiate thinking from CT. This framework brings the common aspects of CT together to form a tool which can be used to navigate the literature; identify relevant resources; structure and develop content; and track the process of CT. This chapter includes some examples of how the framework can be used in education and everyday life. However, to further demonstrate its potential uses, it has been applied throughout the research findings to show how the framework can help pinpoint which aspects of CT educators are focusing on, and additional opportunities for development.

Chapter 3: Educator Perceptions of Critical Thinking – This chapter contains the methods and findings from the educator perception survey and addresses Research Question 1. This first cross-disciplinary, multi-institutional survey relating to CT in Australia was designed to capture educator understanding and intentions for CT development and assessment in their classrooms. It describes the rationale, and the construction and analysis of the educator perception survey approach used to investigate educator understanding of CT. In addition to presenting an Australian perspective on CT, this chapter also demonstrates how multivariate analysis approaches, typically used in ecological studies, can be applied to survey data to enable complex hypothesis testing and facilitates deeper understanding about survey responses enabling the generation of educator profiles.

Chapter 4: Exploring CT Development in Chemistry – This chapter begins to address Research Question 2. It presents the case study methods and the first of the case study findings. The chapter includes a mini literature review on CT assessment and explains the tool selection and case study method used to quantify educator actions and corresponding student outcomes. It also describes the CT development outcomes from studying a semester of first-year chemistry, where the teaching strategies included a fortnightly workshop (which incorporated a range of activities including quizzes, worked question sets, group work and novel case study problems) in conjunction with traditional lectures and traditional laboratory experiences.

Chapter 5: Exploring CT Development in Biology and Society – This chapter reveals the second of the case study findings. It explores the outcomes of studying a biology-based science and society topic. This course incorporates elements such as guest lecturers, videos, popular press, and scientific articles into weekly class discussions and lectures. This was a two-part case study. The first round of data collection replicated the method used in the chemistry case study – gathering information about the existing approaches used in the course and exploring their effects on student CT development. However, the second round of data collection incorporated interventions to target CT skills that were identified as areas of weakness in round one.

Chapter 6: Comparing CT Development in Science - This chapter continues to build understanding for answering Research Question 2. It reflects on the student's perceptions, observed actions of educators, and compares the CT performance changes from the biology and chemistry case studies to explore if the way time is spent in the classroom produces a detectable difference on CT development.

Conclusion: From Concept to Classroom – The conclusion revisits all key study findings and highlights the original contributions of knowledge in this thesis. It offers a perspective on the intentions and actions of Australian science educators and suggests future research possibilities and analysis approaches.

My chapters (and primers) address the goals of this transdisciplinary research by each adding to the collective understanding about CT and its development. The data chapters have a sole focus on Australian educators to contribute understanding to an area where knowledge is currently sparse. In addition, all data chapters employ methods from science, psychology and education to assist with generating **usable** knowledge about CT. The findings of this research are intended to inform science education; however, the broader understanding about CT and the methodology are applicable to all educators. In terms of specific contributions, my ACT Framework is an original contribution that synthesises the ideas about CT to help overcome the issues around defining CT that were raised in the literature review (addressing aim 1). The CT framework is also applied throughout the research to show the different ways it can address gaps in educator understanding about CT. This framework was developed with educators (my broader stakeholders) in mind, wanting to supply them with an approach for understanding research on CT. The following data chapters then involve my key stakeholders – science educators. The first of the data chapter captures the ideas held by these stakeholders (addressing aim 2). Then the proceeding data chapters further explore these perspectives around teaching approaches that science educators use to development CT (addressing aim 3). These case studies were designed to increase awareness about teaching techniques in science that are effective for developing CT to assist science educators with making decisions about CT development in their classrooms. This research compares the effects of different CT development approaches in physical and life sciences to see if any lead to stronger CT development outcomes. This research configures an approach that educators can adapt to their own classroom. It demonstrates how to assess CT using observational tools, as well as general, qualitative and skill-specific assessment instruments. The body of work serves as a tested model for assessing the development of CT. Educators can implement some (or all) of these evaluation approaches in their own classrooms. Its contribution is not just to science; it is an adaptable package which models development and evaluation approaches that are relevant across disciplines and education levels (even though the tools will change based on age-appropriateness).

CT continues to be emphasised as an important skill in Australian education. In a recent publication, Oliver and Jorre de St Jorre (2018) reflected on the qualities Australian graduates currently hold and will need. In their analysis, they found that CT and global citizenship were listed as graduate attributes for 87% of the education providers they examined. However, when it came to the Student Experience Survey results, there was a 7-10% difference in student perception of their knowledge of their field and their CT skills, irrespective of their point in their degree. This highlights a disparity in the way our education system demonstrates the value and relationship between content knowledge and CT skills. This is not surprising given that content knowledge is made explicit, and CT skills are embedded, but it is problematic if we want Australian students to succeed into what Gross (2017) describes as a post-truth world. In fact, Arum et al (2012) found that graduates possessing underdeveloped CT skills experienced poorer outcomes. These include three-times higher unemployment rates, greater credit card debt, and taking longer to cohabitate or marry than higher-performing peers. A growing picture of the importance of CT in modern society is developing, but difficulties can arise when trying to give practical directions for bringing it about. Oliver and Jorre de St Jorre (2018) recommended that to achieve the strongest outcomes, graduate attributes need to be communicated and explained repeatedly, and assessment needs to be explicit. These recommendations are not dissimilar from those made by the Delphi panel in 1990 (Facione, 1990a, 1990b), but it is a shift away from the way things are currently done in Australian education. Science education researchers need to do more to develop CT in our students. This means we need to understand more about the current state of CT development in Australian classrooms.

PRIMER 1

Scientists make sense of the world around them. We deal with theoretical problems by making observations, undertaking experiments and challenging knowledge until we clarify the uncertainty. We compare things that we do not know, against the things we do know, to see what theories hold. We follow the so-called scientific method and put our faith in numbers to help us make our judgements. But what a non-scientist might not consider is that we experiment bravely. We recognise processes do not always work out the way one might expect, but we proceed anyway. We understand that sometimes the technology to which we have access is not yet advanced enough to answer our question completely, but we proceed anyway. We recognise that the existing theories that we trust today, the very ones that underpin our research, may need to change with the knowledge we discover tomorrow. But we proceed anyway because we seek to evolve knowledge.

Information on CT is not lacking, but clarity is. Throughout the upcoming literature review, I reflect upon the multifaceted nature of CT and the need for a better way to consider it. This is followed by a chapter which tackles this. The Adaptive Critical Thinking (ACT) Framework and definition were a result of a series of thought experiments – considering what CT is, and what it is not. It arose out of wanting to bring a scientific approach to a theoretical problem. In evolutionary biology, we work by the theory of parsimony – that is, the simplest solution is probably the right one (at least until we are faced with new evidence). However, before arriving at this state we collect and compare evidence about the organisms we are studying and its close relatives. We consider what it is, and what it is not. We seek to characterise. In biology, this process of recognising and describing organisms is known as taxonomy. As I synthesised the literature, I applied this descriptive classification approach to CT - moving from discussing the similarities and differences in perspectives on CT to evolving understanding about CT to a more universal, yet adaptable stance on CT that I have called the ACT Framework.

Chapter 1

Literature Review

The greatest obstacle to discovery is not ignorance, it is the illusion of knowledge.

Boorstin, 1983, p.86

1.1 Introduction and chapter outline

Existing reviews on critical thinking (CT) serve to describe the skills, dispositions and abilities associated with CT (see Facione, 1990a; Kennedy, Fisher and Ennis, 1991; Paul et al., 1997b; Pithers and Soden, 2000; Moon, 2004; Lai, 2011a; Sanders and Moulenbelt, 2011; Niu, Behar-Horestein and Garvan, 2013; Thomas and Lok, 2015; Davies and Barnett, 2015). This review examines the shift in ideas about CT both prior to and following on from the Delphi panel's attempt to generate a consensus view (Facione, 1990b). This review explores why despite the efforts of Facione (1990a, 1990b) and the others who followed (Paul and Elder, 2006a; Halpern, 2003; Paul, 2011, 2012; Ennis, 2011a), CT seems to be "less, rather than more clearly defined" (Capossela, 1998, p.1). It explores the impact of Facione's attempted consensus on CT (Facione, 1990a, 1990b) and builds on Facione's recommendations (1990a).

This literature review provides an exposition of the evolution of ideas about CT in modern education and highlights that new thinking on CT has reached a saturation point. There seem to be too many perspectives – each with their own merits, yet none which fully capture all the parts (Johnson and Hamby, 2015). Yet by considering the various perspectives collectively, meaningful insights emerge into the actions involved in CT, the standards for good CT, and the qualities of a critical thinker. There is considerable agreement about core skills and dispositions associated with CT (Sternberg, 1986; Lai, 2011a), with differences evolving from the recipe for putting them together. This chapter demonstrates that there is indeed still a need for a shared understanding about CT and posits that a framework, like Tarricone's (2011) work on metacognition is needed to bring cohesion to the thinking on CT. It also addresses Johnson and Hamby's (2015) recommendation to acknowledge extant definitions prior to redefining CT. I pay particular attention to barriers to consensus and the features and failures of current CT frameworks. This sets the scene for my solution: the Adaptive Critical Thinking Framework (the bulk of which is presented in Chapter 2). This framework initially emerged out of my own synthesis of the literature, as the plethora of ideas signalled a need for a common set of terms to help educators navigate through the various perspectives on CT and incorporate the contextually relevant ones into their classroom. It is a tool that can be used to bring consistency to the field, increasing the usability of ideas about CT among philosophers, psychologists and educators.

1.2 Definitions of CT

CT has been defined both broadly and narrowly (Kennedy et al., 1991). Early ideas about CT from Dewey (1910) and Glaser (1941) address the nature of CT, and introduce components such as disposition, knowledge, skill and reflection. Bloom (1956) explored the stages of thinking and proposed a hierarchy of educational objectives. Ennis (1962) contributed a concept of CT, incorporating features of existing theories, and adding in his own ideas about the "correct assessing of statements" (p.83) which shifted the focus of ideas onto knowing and applying criteria to judgments. Discussions about the specificity of criteria (domain-specific or general), as well as skills and dispositions associated with CT, increased through the 1980's. Yet despite the explosion of perspectives over that period, questions remained about the motivation for CT, the dimensions that make up CT and the criteria for assessing CT. It was around this time that Facione (1990a, 1990b, 1990c) convened a panel of 46 educators from Canada and the US (referred to throughout as the Delphi panel) with the intent to discuss CT and generate a consensus about its development in order to generate a collective approach for education.

The Delphi panel noted the multifaceted nature of CT, acknowledging components including cognitive skills (see Table 1.1), cognitive dispositions and affective dispositions (see Table 1.2). The definition that emerged from their deliberation was "purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based" (Facione, 1990a, p.3). In short, CT involves habitually employing thinking skills to reflect, evaluate and explain with clarity. The Delphi process revealed that CT is a convoluted construct, and whilst a consensus view was generated, Facione was quick to note that not all the experts agreed with all the ideas put forward. Some of these experts went on to generate and/or revise their own conceptions of CT (Ennis, 1991; Ennis, 2011a; Facione, 2000; Facione, 2015; Lipman, 1995a; Lipman, 2003; Paul, 1996; Paul, 2005; Paul, 2011, 2012), and continue to debate about the components and meaning of CT.

Table 1.1 Summary of the Delphi panel's consensus ideas regarding the cognitive skills and sub-skills associated with CT. ***Note.*** **Adapted from 'Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction' (Facione, 1990b, p.12-19).**

Removed due to copyright restriction

Table 1.2 Summary of the affective components from the Delphi panel's consensus ideas regarding CT. ***Note*****.** **Adapted from 'Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction' (Facione, 1990b, p.25).**

Removed due to copyright restriction

The definition generated by the Delphi panel is widely used, but like many of the perspectives before it, the full body of the work seems insufficiently explored. It seems to have become just another perspective in the ongoing discourse. Reflecting on the past 26 years since the publication of the Delphi panel's review of CT (Facione, 1990a, 1990b) ideas about CT have continued to develop. For example, while current perspectives on the cognitive skills associated with CT still generally align to the Delphi panel's 1990 conception, ideas about metacognition and creativity provide some areas of discrepancy. With some believing CT is a metacognitive process (Dwyer et al., 2014); and others adding divergent thinking (fluency, flexibility, originality and elaboration of ideas; see Vincent, Decker and Mumford, 2002) and creativity into the mix of skills (Fisher, 2001 and 2011; Almedia and Franco, 2011). A number of frameworks about CT have also been developed (see Paul and Elder, 2006a; Beghetto, 2002; Duron, Limbach and Waugh, 2006; Edwards, 2007; Rabu, Aris, and Tasir, 2013; Dwyer et al., 2014). However, of these, it is the Paul-Elder framework that has been most widely promoted and adopted.

The extent to which CT abilities are general or domain-specific has remained an ongoing node of discussion (McPeck, 1990; Ennis, 1991 and 2011b; Possin, 2008; Robinson, 2011; Vainikainen, Hautamäki, Hotulainen, and Kupiainen, 2015), but there has been a shift towards incorporating the learning opportunities from both perspectives (see Sections 3.2.2-3.2.4.). However, some additional conceptions of CT have been put forward, including – a dispositional theory of thinking (Perkins, Jay and Tishman, 1993); the "honest evaluating of alternatives" (Hatcher, 2000, p.6); a practice of re-evaluating one's thinking and reasoning (Possin, 2002); a notion of self-critical thinkers (Andrews, 2009); and thinking that recurrently accesses and restructures knowledge (Almedia and Franco, 2011).

Habits, attitudes, behaviours and dispositions are still included in conceptions of CT (see Facione, 1990a, 1990b; Paul, 1992; Kuhn, 1999; Halpern, 2003; Paul, 2005; Ennis, 2011a, 2011b; Hatcher, 2013; Ennis, 2018). Understanding of dispositional and motivational aspects have increased (see Halpern, 1999; Pithers and Soden, 2000; Macpherson and Stanovich, 2007; Ku and Ho, 2010; Ennis, 2011b; Sosu, 2012; Miele and Wigfield, 2014; Facione, 2015). The disposition to think critically is now distinguished from the ability to think critically (Ennis, 1985; Facione, 2000; Kuhn, 2019), and this increases possibilities about how CT might be targeted in the classroom (Bloch and Spataro, 2014; Kuhn, 2019). However, there is still some debate about the role of dispositions (Schmaltz, Jansen and Wenckowski, 2017) and factors for predicting CT performance are still under exploration (see Clifford, Boufal and Kurtz, 2004; Celuch, Kozlenkova and Black, 2010; Manalo and Sheppard, 2016).

Facione's work with the Delphi panel led to the consolidation of the existing thinking on CT through the process of attempting to provide an operational definition to encourage its development and assessment in the classroom. However, ideas about CT have continued to develop as theorists and educators gather empirical data on thinking skills, dispositions and metacognition. Given the reflective and evidence-driven nature of CT, this ongoing adjustment to the construct is conceivably expected and appropriate. But rather than adjusting and improving the construct, theorists appear to be stuck trying to re-invent aspects which other fields have already described. The various perspectives and possible causes driving the reinvention of similar ideas are explored below in Sections 1.2.1-1.2.3.

1.2.1 Is there a consensus on any aspects of CT?

Views about CT are expanding and transforming. For example, Paul initially proclaimed that CT could not be reduced into a single definition (Paul, Binker, Martin, and Adamson, 1989), but he later contributed a perspective that CT is 'thinking about thinking to improve it' (Paul, Fisher and Nosich, 1993). Ennis has also revisited his definitions of CT - initially "the correct assessing of statements" (Ennis, 1962, p.8) to update it to include "reasonable, reflective thinking" (Ennis, 1991, p.6) acknowledging that CT is responsive to context and its transformative power lies in using CT skills to refine thoughts about "what to do or believe." Facione's work with the Delphi panel was the first major attempt to consolidate education-focused philosophies about CT (Facione 1990a). However, as described above, additional ideas have been formulated since this initial 'consensus' view. As a result of the ongoing discourse, there is a pervading belief that a suitable definition of CT is yet to be fashioned (see Poirier and Hocker, 1993; Sanders and Moulenbelt, 2011; Lai, 2011a; Moore, 2013). Johnson and Hamby summarise the state of a consensus on CT as "an overabundance of problematic definitions" (2015, p.1), resulting from a failure to properly acknowledge existing definitions and identify their strengths and weaknesses.

Ideas about CT have shifted with the changing demands on society, which now require individuals to successfully navigate through a surplus of information that has varying degrees of validity. The focus has changed from seeing CT as it functions in education, to recognising CT's role in the workplace (for example Gambrill, 2005; Prinsley and Baranyai, 2015; Hirsh, 2018) and as a life skill (Pellegrino and Hilton, 2012). Even though there is still no unanimously accepted description of CT, there is acknowledgement that CT incorporates many thinking processes such as problem-solving, inquiry, reflection, interpretation, justification, planning, argument analysis, argument construction, decision making, appraising assumptions, reasoning and self-control (Lai, 2011a). Further, many definitions consider reflective and/or reasoned judgment to be the major outcome of CT or being a critical thinker (Dewey, 1910; Ennis, 1991; King and Kitchener, 1994). These similarities show that there is clarity about certain aspects of CT, so it is conceivable that a consensus stance could be achieved. Yet, there also seem to be factors holding up the process. For example, the multifaceted nature of CT and its application make it challenging to achieve a consensus approach and stance. What is needed is a meta-approach, such as an adaptable framework definition that facilitates the separation of the components (general in nature) and their applications (context-dependent) to help users understand which aspects of CT theory they are drawing from. As will be discussed in section 1.2.3 there are existing CT frameworks; however, they do not acknowledge the different branches of ideas about CT while relaying the degree of similarity and connectedness between core elements. A framework can be successful at amalgamating various ideas about CT (see Chapter 2), but the factors driving the ongoing generation of ideas about CT must be addressed to enable the definitions to be unified. These barriers are explored next.

1.2.2 Sources and barriers to achieving a consensus view

CT is a thinking process that is applicable and valued across disciplines. However, there is yet to be a stance that captures the variability. This section will discuss how the barriers preventing the generation of a consensus stance are a product of the multifaceted nature of CT. Essentially these barriers relate to a failure to create or use a shared language, a failure to generate a perspective that is sensitive and adaptable to context and a failure to synthesise and build from existing beliefs about CT across discipline boundaries.

The ongoing generation of ideas about CT seems to be a self-perpetuating barrier to achieving a shared understanding. The range of perspectives on CT arises from the fact that they are formulated from three different conceptual traditions - philosophy, psychology and education (Sternberg, 1986). The philosophical stance on CT has a set view about the **characteristics** a critical thinker possesses (Lewis and Smith, 1993), with an emphasis on logic (Sternberg, 1986), “perfections of thought” (Paul, 1992) and qualities and standards (Bailin, 2002). The psychological stance has a more behavioural-based focus when it comes to CT (Lai, 2011a). Theorists from this tradition are concerned with both the **attributes** (skills and dispositions) and **actions** of a critical thinker, as well as the everyday constraints that might influence the development or deployment of these attributes (Sternberg, 1986). The educational stance traditionally focuses on the development of **cognitive skills** (Bloom, Englehart, Furst, Hill and Krathwohl, 1956; Anderson, Krathwohl and Bloom, 2001). However, this conceptual tradition is now shifting to include dispositional aspects, since these can lead to better development outcomes (Bloch and Spataro, 2014). Having different fields contributing to CT definitions is not problematic in itself, provided that there is shared language across the disciplines. However, what is evident from the existing literature is that there is not a common set of terms, and the academics from differing fields often frame their views in contrast with each other.

While CT definitions are nested in three conceptual traditions, the perspectives about CT are not that different. Sternberg (1986) pointed out that the main differences in the definitions of CT seem to emerge when it comes to the application of CT skills and dispositions, rather than the skills themselves – particularly with regard to the prompt/drive, context and proficiency of their use. Related to this is the setting for CT. Sanders and Moulenbelt (2011) describe the main distinctions between definitions as clustered into "context-specific definitions" in which CT "cannot occur without a specific context" (p.44); and "cross-disciplinary definitions" in which CT "skills are not dependent on a particular context" (p.45). These two different stances on the role of context are one of the driving sources behind the hold up to a consensus stance. Additionally, it is likely that a hold up to arriving at a consensus is that existing perspectives are not capable of the nuanced applications of the core skills.

The multifaceted nature of CT and its applications make it challenging to achieve a consensus stance. As a process, CT encourages individuals to consider alternatives, weigh up options and recognise how different sets of circumstances may change the best response to a scenario. However, the specific parameters for this process and rules guiding the quality of the outcome are shaped by context. Trying to create a definition that describes the actions of CT incorporating both general and domain-specific constraints could also be responsible for thwarting efforts to achieve consensus. Yet is clear from the literature that there are domain-specific ways of knowing (Weinstock, Kienhues, Feucht and Ryan, 2017), despite the general value of CT to any and all thinking and judgements. It is these epistemic details that provide nuances to how CT should look and proceed in different contexts. This means the nature of CT is its own barrier to achieving a shared stance on CT because the specific cognitive demands needed to do CT at any particular time will change depending on the contextual nuances and criteriological parameters of the scenario.

There is an inherent contradiction in attempting to reduce this complex thinking process into a prescriptive perspective. Yet not having a shared-understanding leads to inconsistencies in the development and assessment of CT and propels further theorising and idea generation. Therefore, the pragmatic need for a consistent approach for exploring CT might outweigh the theoretical conflicts about a reductive approach to multifaceted nature of CT. The main barriers to reaching a consensus arise from attempting to construct a 'one-size-fits-all' definition which does not accommodate these variations in application nor clarify the differences in terminology. In fact it is where and how these ideas are being generated that is driving the real issues preventing a consensus approach. Issues relating to language (with regard to terminology usage and the communication of ideas) and the alignment and relevance of the theories to practice are covered next.

Terminology barriers

Terminology is a major barrier to achieving a consensual view. Each iteration of definition generation incorporates some of the existing language in a slightly different way yet does not always make it clear which components of CT have shaped the view. Sanders and Moulenbelt (2011) noted, "there is no shortage of scholarship on critical thinking," however, each contribution tends to discuss CT as if "their personal or discipline-specific definitions are consistently shared by all" (p.38). Further amplifying this issue is the fact that terms which have etymological nuances are used interchangeably (Cuban, 1984); or sometimes just incorrectly and in doing so are completely changing their meaning, such as Costandius et al. (in Davies and Barnett (2015) misquoting Ennis' perspective on CT as "reflexive thinking" (p.547) rather than "reflective thinking" (Ennis, 1991, p.6). In the case of Costandius et al. (2015), the use of the term "reflexive" (automated and reactive) construes CT as a bottom-up rather than the top-down thinking process implied by "reflective" (in terms of actively monitoring thinking), which is a major conceptual shift in ideas about CT.

Another language issue emerges from how the terminology is used, particularly when the same words are used to convey different meanings. For example, when the Delphi panellists were deliberating over the components of CT, there was a common inclusion of a notion about 'good thinking.' However, when Facione (1990a) reviewed this idea and asked further questions, he ascertained that there were differences in the meaning of ‘good’ when applied to CT by the different panellists. Sometimes it referred to the quality of thought, other times to the ethics and morality of thought. In other instances, good thinking referred to the skilled and habitual application of CT skills, not just the use of the skills themselves. The use of ‘good’ in each of these descriptions demonstrates how judgments about CT are sensitive to context and criteria, and how examining CT through different lenses (skills versus disposition versus morality) can lead to different conclusions about what qualifies as ‘good’ quality thinking. The scenario above highlights how/why seemingly different definitions have emerged from the three fields (philosophy, psychology and education) because they have used different lenses (criteria) when describing CT, shifting the focus of the meaning of CT from expectations (standards - philosophical), to ideals (proficiency - psychological), to actions (skills and products - educational). The disparity is further compounded by the fact that many theorists try and put forward a view that seems mutually exclusive despite the fact that they are situated in a common theoretical tradition (Johnson and Hamby, 2015).

Barriers caused by the communication of ideas

Hindrances in achieving a consensus are not limited to the terminology used to describe CT and its components. Barriers are also linked to the communication of ideas. Confusion and assumptions about CT can arise if a theorist does not explain if the definition or component of CT they are presenting as either essential for CT, or non-essential, yet something that enhances the CT process. Facione (1990a) encountered this problem when unpacking ideas about the dispositions of a critical thinker. Two-thirds of the Delphi panel felt both cognitive dispositions and affective dispositions were core to the meaning of CT, whereas the rest felt the affective components were a bonus and not central to CT itself. In both instances, the affective components were still seen as related to CT, but not necessarily the essence of CT. However, these complexities are not automatically communicated through the term ‘disposition’ nor through a list of components CT includes.

Communication about the extent of inclusion of a particular skill within perspectives is also important. It can reveal the contextual and criteriological differences between new and existing definitions and prevent the misrepresentation of ideas due to hidden nuances in the way the terms are used. For example, CT is sometimes discussed as a series of standards to hold thinking up to, yet other definitions focus on CT as a process involving different skills whose deployment may be influenced and/or complemented by personality traits and dispositions. Kuncel (2011) explains that there also seem to be two different versions of CT skills – one that covers domain-specific skills which develop with practice and expertise (not readily generalisable); and the other which considers a logic-focused "finite set of very specific reasoning skills" (p.2) which could be generalisable (even though evidence suggests they do not always transfer). Again, knowing which version or nuances theorists are working from is important to help achieve a shared understanding, even if this 'understanding' is that perspectives are different.

However, a further communication issue arises because CT is discussed in both a noun and a verb form. Moon (2007, p.126) notes that "the thinking and representation of thinking are different activities" yet CT is a term "that tends to be used to cover both the mental activity" and its product. Sometimes both forms are used within a single definition, making what CT is versus how it occurs/proceeds difficult to grasp. Anderson et al. (2001), helped address this issue for one of the perspectives when revising Bloom's 1956 taxonomy, by shifting the language from noun to verb form to help educators implement them. Yet many other theorists have not further clarified their own or others' perspectives. When definitions, terminology and/or related expectations remain unclarified, they fuel the ongoing confusion about CT, by leaving perspectives open to interpretation even though they may have been formed from very specific ideas. With all these differences in meaning, it is likely that educators would find it challenging to choose which ideas about CT should guide their practice. However, there is a further issue for educators - the gap between theory and classroom.

Barrier caused by the disconnect between theory and practice

A further barrier to the generation of a consensus view is that the theories do not always align with the contexts in which they are used. This could be a consequence of the language and value issues described earlier, however, Norris (1992) explains that there is still a tendency to treat the concept of CT philosophically and in an abstract manner, and as a consequence the resulting definitions do not necessarily reflect any actual reality. For example, to help combat the generation of abstract definitions which might be disconnected from the education settings in which they are used, recent literature has offered a classroom-based approach to interpreting CT. For example, Moore (2013) found that a multi-dimensional view of CT emerged when he explored how the idea of CT is used in educational practice at an Australian university. Moore (2013) highlighted this disconnect, by noting the distinction between his findings and the 'CT movements' attempts to produce a more singular and readily identifiable cognitive mode or perspective in their approaches. However, Moore's investigation was limited in that it only explored a few opinions, of a few closely related disciplines (history, philosophy and cultural studies) at one university. So whilst it is insightful that educational practice in some of the humanities disciplines at this university seemed to favour a multi-dimensional approach to CT, this study does not give a broad perspective of the role of CT in tertiary education (and cannot even give clarity among the three disciplines presented). Further, the potential for application of Moore's (2013) findings is limited, as there is no empirical evidence provided to show that the academic's opinions of important attributes they sought to encourage in their students were apparent in their teaching. However, the findings show that the classroom-based insights educators have of CT add depth to the understanding of CT theory.

Norris' (1992) ideas about the abstract nature of CT definitions build on Wittgenstein's (1958) thoughts about deriving a word's meaning by examining situations where the word is being used. For example, Wittgenstein discusses that it can be difficult to verbalise what the colour 'red' is without examples of things that are red and things that are not red. Therefore with definitions of CT, it is important to look to examples of CT and then discern from it, to reveal how to distinguish 'sound and reasoned judgment' from other cognitive processes such as thinking or judgment (which, on their own, are not CT). Further, a scientist, artist and philosopher could all have a very different perception of what red is and they may have a way of describing it that picks up on unique or discipline relevant attributes. Similarly, ideas about CT are also affected by context. The traits, skills and dispositions associated with CT are valued or weighted in different ways depending on the theorist's background (philosophy, psychology, or education) and the domain-specific needs of the situation (criteria for sound reasoning in art versus science versus every-day life etc.). This perception variation is reflected in the variety of definitions for CT. Yet instead of acknowledging crossover between views (and that new ideas may be discipline-specific insights into an existing definition), theorists appear to use alternate phrasing about CT, whilst critiquing existing ideas, to unveil their insights as novel conceptions. Ironically this has created an ongoing series of opposing ideas about CT in the pursuit to generate one unified definition.

By concentrating on the abstract examples and differences rather than the similarities between perspectives, theorists have cluttered non-experts understanding of what CT is. The reality is the different epistemologies held by each discipline mean there are context-specific nuances to the application of CT. These differences result in the need for different sets of thinking resources to arrive at sound judgments in each discipline. Paul (1996) notes that these "insights from multiple disciplines (without losing coherence or rigor)" (p.34) are a core part of what is missing from past research into CT. This indicates that no single view about CT is necessarily a 'more correct' reflection of reality than another – they are all "imperfect" (Paul, 1996, p.34). However, there is value in exploring the various expressions of CT through the different lenses in these discipline contexts, to help clarify what is CT and what it is not. Further, a common command of the term is needed (Johnson, 1996), and "more effective" communication between theorists and "those concerned with classroom instruction" is required to be able to reintroduce "the art of thinking critically…in education" (Paul, 1996, p.34). These requests and recommendations eventually resulted in the emergence of a number of frameworks, which attempted to help mobilise CT definitions and clarify the components and process of CT. However, as discussed next, a widely applicable approach is still missing and a solution such as an adaptable framework approach is needed (see Chapter 2).

1.2.3 The features and failures of previous frameworks

As presented and discussed throughout this chapter, numerous attempts have been made to arrive at a common state of belief about CT. Yet through 25 years of debates and discussion, the literature on CT continues to supply new opinions and perspectives on CT and its development (Dwyer et al., 2014; Kwan and Wong, 2015; McCormick et al., 2015; Huber and Kuncel, 2016; Vainikainen et al., 2015; Ennis, 2018). However, this ongoing generation of information is counter-intuitive to successfully implementing these ideas into education. Particularly as educator perspective studies have revealed educators do not have time, nor resources, nor the training to be able to research and implement these tactics in the classroom (Paul et al., 1997a, 1997b; Shell, 2001; Black, 2009; Choy and Cheah, 2009; Aliakbari and Sadeghdaghighi, 2012; Kowalczyk et al., 2012; Stedman and Adams, 2012; Moore, 2013; Alwadai, 2014). These perception studies highlight that educators are not in need of more information; instead, they need a strategy to help them navigate through the existing information and incorporate the relevant components into their pedagogy.

There are a number of existing frameworks on CT, which educators could use as a starting point. Some of these outline aspects of CT (Paul and Elder, 2006a; Van Gyn and Ford, 2006; Dwyer et al., 2014; Thomas and Lok, 2015). Other scholars focus on evaluating teaching approaches to encourage and improve CT development (Duron et al., 2006; Rabu et al., 2013; Osborne, Kriese, Tobey and Johnson, 2014); or use CT to investigate other constructs (Beghetto, 2002). There is also a framework which is a mix of the concept, development and evaluation (Edwards, 2007). A critique of each of these follows.

The first framework educators could consider concerns a construct of CT by Paul and Elder (2006b). This framework is available through the Foundation for Critical Thinking website. Educators may also find this website useful because it has a number of resources for CT development. Their main framework, developed by Paul and Elder (2006b), involves elements of thought, standards of excellence and intellectual traits (See Figure 1.1). The Paulian definition is framed around “the art of analyzing and evaluating thinking with a view to improving it” (Paul and Elder, 2006a, p.2). It is a generic approach that can be applied across “all subjects,” and is heavily embedded in the philosophical stance with a focus on perfections of thought. As such, the application of the framework is directed at overcoming “native egocentrism and sociocentrism” through the “mindful command” of “standards of excellence” (Paul and Elder, 2006a, p.2). The components included in this framework are logical and incorporate many of the important aspects of CT. However, the main criticism I have of this perspective is that their purpose of CT assumes that everyone should be focused on improving their thinking. Yet not everyone’s interest is academically based, meaning the value of CT embodied by this approach may not be as evident to, nor seen as applicable by, these individuals. While the 3-part framework captures the underlying essence of being a critical thinker, it is also a little deficient, in that there is more to the purpose of doing CT than improving thinking. For example, King, Goodson and Rohani (1998) summarised some perspectives in CT, noting it can be approached in terms of problem-solving; higher-order thinking; “results produced by thinking creatively” (p17); metacognition; and dispositions. Each of these flavours to CT will shift the purpose and also how the process is undertaken. King et al. (1998) also draws attention to the fact that the “successful application” of thinking skills are “explanations, decisions, performances, and products” (p.1). The outcome is not always focused on improving thinking.

Removed due to copyright restriction

Figure 1.1 Aspects of the Paul-Elder CT framework. ***Note.*** **Reprinted from 'The Miniature Guide to Critical Thinking: Concepts and Tools' (Paul and Elder, 2006b, p.21).**

This deficit does not render the Paul-Elder framework invalid, because the aspiration to improve one's thinking is in the spirit of being a critical thinker. The framework also helps emphasise how CT the process benefits from being explored as a union of metacognitive skills, dispositions and thinking skills. It is these elements which help ensure high standards for the thinking outcome – another worthwhile goal for educators. However, noting the deficit, serves as a caution for educators wanting to use the Paul-Elder approach. Instead of making it a sole stance on CT in their classrooms, educators may be better served by treating it as a standard of thinking for students to aspire to. Lastly, education systems that employ rote-learning approaches and recall-style assessments are not structured to convey the value of good thinking to students, so this framework would not be applicable in these settings.

Overall, the Paul-Elder framework is a valuable tool, because it creates a way to define the quality of thinking, expectations around reasoning and reflection by having these standards of thinking, and intellectual traits to apply to the elements of reasoning. However, users of the framework need to be aware that by focusing on the reasoned and reflective aspects of what good judgement looks like, this approach has lost sight of the core judgement-based purpose of CT. If educators are having difficulty encouraging students to engage in the idea of thinking well, perhaps they can look to other definitions to help generate student motivation about the value and purpose of CT.

Van Gyn and Ford (2006) reconceptualised the Paul and Elder approach. Their definition maintains an emphasis on the quality of thinking, however, it also characterises the types of thinking tasks and criteria that are relevant to CT. Educators may find this definition more user-friendly than the Paulian approach. Van Gyn and Ford (2006) have also constructed a guide intended for tertiary educators, which educators searching for resources on CT will find helpful, because it not only contains definitions and frameworks, but it also has worked examples in different domain areas. For the purpose of reviewing their framework, I present the image containing the culmination of their ideas for the instructional design process (See Figure 1.2). It includes their core elements (intellectual habits, intellectual deliberation and reflexive dimensions) as well as instructional design components. Van Gyn and Ford (2006) have assembled the framework this way to make the aspects and interactions overt, with the hope that its explicit nature will make it easier for teachers to use. This framework is useful because it shows how having a known CT definition leads to clearer structure for planning instructional strategies, guidelines and activities for students, as well as assessment goals. However, the main limitation of this model is it is so explicit that it does not give educators the opportunity to vary components should they have a different view on that aspect of CT.

Removed due to copyright restriction

Figure 1.2 Van Gyn and Ford's instructional design framework, highlighting the relationships between their definition, dimensions and criteria for CT as well as the suggested standards for applying them to planning, assessing and evaluating student work. *Note.* Reprinted from 'Teaching for Critical Thinking' (Van Gyn and Ford, 2006, p.34).

Dwyer et al. (2014) also have a strongly metacognitive view of CT. They see its role as a "purposeful, reflective judgement" process that "increases the chances of producing a logical conclusion to an argument or solution to a problem" (p.43). However, their framework is comprised of both cognitive and self-regulatory aspects (see Figure 1.3), which they argue "are necessary for the successful application of CT" (p.49). Dwyer et al. present this framework at the end of an extensive review of the literature, where they gather the evidence-based perspectives on various aspects of cognition and executive function to argue their case for the interconnectedness of the elements they use in their framework. Expressing how some attributes guide the application of other attributes (the self-regulatory functions guiding the use of CT skills such as evaluation and inference), while things like reflective judgement and CT are correlated and develop in an "interdependent, cyclical manner" (p.48).

Removed due to copyright restriction

Figure 1.3 Dwyer, Hogan and Stewart's integrative cognitive framework of CT. ***Note*****. Reprinted from 'An integrated critical thinking framework for the 21st century,' (Dwyer et al., 2014, p.49).**

Dwyer et al. (2014) do little to explain how their framework could be applied, merely suggesting it represents "a cognitive framework of learning outcomes" that can address "the lack of impetus focused on training CT skills" (p.49). Given the framework is comprised of mental processes and skills, it seems less about learning outcomes and more of a construct. Nevertheless, their statements lend itself to the inference that by making the components of CT explicit for educators, that educators can then make these into explicit learning outcomes for students. Again, this is a framework that should not be used in isolation from the supporting article, where users should specifically refer to the component segment (or learning outcome) they intend to focus on to gain insight into the envisioned standards and expectations. As such, educators who have developed their own understanding of CT may find this framework too inflexible for their needs.

Thomas and Loc (2015) supply a framework with a much more general approach to CT than Paul and Elder (2006a), Van Gyn and Ford (2006) or Dwyer et al.(2014). Thomas and Loc developed a conceptual framework that characterises CT as having three interconnected attributes (skills, dispositions and knowledge), with each of these attributes also having three composite sub-skills (see Figure 1.4). Thomas and Loc also define the purpose of CT as "supporting the quality of reasoning and subsequent judgements" (2015, p.101). However, it is initially unclear if they are implying that the conceptual framework attributes are the components that support the 'quality of reasoning and judgement,' because they also propose three performance levels (which represent different things to the conceptual attributes) that offer support to CT. The connection between the framework and performance levels is not made clear till later in the article, where the authors briefly express the need to map "activities and learning outcomes" to the framework themes, and then use these to develop "functional attributes across the performance levels"(p.103).

Removed due to copyright restriction

Figure 1.4 Thomas and Lok's operational framework for teaching CT. *Note*. Reprinted from 'Teaching critical thinking: An operational framework' (Thomas and Lok, 2015, p.98).

The purpose of Thomas and Loc's framework was to both clarify the important parts of CT relevant to education, and also provide components that could be modified based on educator context (learning setting or discipline). The attributes included in the framework are sufficient to allow educators and researchers to examine and categorise literature on CT. However, there is no explanation of the interaction between each of the attributes (implied through the arrows in Figure 1.4) to frame how users might appropriately engage or incorporate multiple attributes and sub-skills in their classrooms. Another aspect that affects the usability of this framework, is the lack of a clear definition to guide its application, or an explanation or set of rules about the supplementation of ideas. The underlying issue is whether the framework could uphold a consistent standard of CT across users. However, Thomas and Loc do define performance with a more simplified perspective of the strongly metacognitive standards around CT than other theorists (Paul and Elder, 2006a; Van Gyn and Ford, 2006). These views on performance, in combination with the framework, provide a tool which educators may find more useful for organising activities around CT and supporting student CT abilities.

The first four frameworks examined have been heavily focused on CT theory. However, educators may be more interested in models that help them to identify opportunities to scaffold CT development. If this is the case, they might like to consider the work by Osborne et al. (2009), who built a framework to assist with the development of CT through interpersonal skills for an online course. Osborne et al.'s framework (See Table 1.3) is based on the ideas of Kuhn (1999), Paul & Elder (2002), Smith (2002) and Doherty, Hansen, and Kaya (2007). By making the theoretical underpinnings explicit, it gives educators without a pre-existing stance on CT a place to start their thinking. Educators with a pre-existing stance on CT can compare their understanding against the explicit explanation to identify the extent of alignment with the theorists, in addition to determining whether the framework will be beneficial for their classrooms. Osborne et al. (2009) also outlines the process they undertook to formulate their ideas and apply them in their classroom. Both types of educators (those with and without an existing stance on CT) can replicate this process for their own setting. Osborne et al. also supplies tools to assess student thinking in line with their model, so this framework (and its supporting materials) provide a full package for educators who are interested in developing both CT and interpersonal skills and also find the framework ideas to be necessary, sufficient and substantive for a perspective on CT. However, an educator may not find this view to be sufficient, because it uses a limited set of CT skills and dispositions, and also strips out some of the explicit and core language around CT which has been found to enhance the efficacy of CT development (Abrami et al., 2008, 2015). Overall Osborne et al.'s (2009) framework provides a highly accessible model for educators who want students to understand how to appropriately challenge their own (and other's thinking) with a more considered approach to logic and argument than structured Socratic questioning (Paul and Elder, 2008; Lee, Kim and Kim, 2014) or identifying logical fallacies (Adler, 1996).

Table 1.3 Summary of Osborne et al.'s CT framework, describing both components and related actions. *The related actions and criteria have been reorganised and on occasion paraphrased to assist an educator's ability to compare this framework against the others in this chapter. *Note*. Adapted from 'Putting It All Together: Incorporating "SoTL Practices" for Teaching Interpersonal and Critical Thinking Skills in an Online Course.' (Osborne et al., 2009, p.47)

Component	Description	Related actions and criteria*
Recitation	"state known facts or opinions"	• Acknowledge what aspect(s) of what is being stated are factual and what are based on opinion.
Exploration	"analyse the roots of those opinions or facts"	• Digging below the surface of what is believed or known. • Working to discover the elements that have combined to result in that fact or that opinion. • Analysis without an attempt to comprehend the impact of those facts or opinions.
Understanding	"involves an awareness of other views and a comprehension of the difference(s) between one's own opinion (and the facts or other opinions upon which that opinion is based) and the opinions of others."	To truly "understand" our own opinion in relationship to others, we must understand how to discover the roots of the opinions of others • Become aware of the roots of our own opinions, • Initiate an active dialogue with the other person about his or her opinions and the roots of those opinions.
Appreciation	"a full awareness of the differences between our views and opinions and those of others."	To truly appreciate differences, we must be aware of the nature of those differences. • Undertake an analysis of the opinion as recited by the other. • Generate a complete awareness of the similarities and differences between our own opinions (and the roots of those opinions) and those of the other. • Be aware that while opinions may differ, we are now in a position to truly appreciate and value those differences.

Another example of a framework for an online setting was developed Rabu et al. (2013), who proposed a framework to describe the different ways that educators' scaffold and educate students in an online learning environment. This framework structures a way to explore student CT engagement, cognitive performance and general CT skills through Nonaka and Takeuchi's (1995) Socialisation, Externalisation, Combination and Internalisation model. Similar to the Osborne et al approach, Rabu et al.'s framework may be helpful for educators, because it explains which perspectives on CT it incorporates, which can serve as a model of how to incorporate existing literature for other educators. However, the main disadvantage of this framework is that it very complex, and therefore may not be readily understood or applied by an educator unless they are already working in an online learning environment and using the Asynchronous Online Discussion Forum to develop CT.

If educators find development frameworks such as Rabu et al. (2013) too convoluted, then they may prefer something that is more generally applicable like the 5-step framework created by Duron et al. (2006). This cyclical framework (see Figure 1.5) scaffolds a process of pedagogical improvement designed to assist educators in developing students CT skills. The framework involves teachers identifying or generating learning objectives; teaching through questioning and discussion; considering and including the types of activities that "promote active learning" (p.162); collecting feedback and documenting participation and progress in tasks; and providing "thoughtful and purposeful" feedback to students (p.163).

Removed due to copyright restriction

Figure 1.5 Duron, Limbach and Waugh's 5-Step Model to Move Students toward CT. *Note.* Adapted from 'Critical Thinking Framework for Any Discipline' (Duron et al., 2006, p.161).

Duron et al.'s framework aims at improving pedagogy for CT by outlining the steps educators should take to promote the development of CT. It uses a skills-based definition of CT that is grounded in Bloom's taxonomy to judge when CT is occurring. However, the framework moves educators through some of the more self-regulatory aspects of CT to help them generate understanding about the opportunities for improvement in their classroom. The drawback for the students is that these self-regulatory behaviours (which theorists such as Paul (1996) believe are important to improve the standard of thinking are not modelled for the students, or part of the planned abilities to give them opportunities to develop in. Although the educator thinking process incorporates a more holistic approach to CT and will be rigorous, the lack of suggestions around modelling and incorporating standards and dispositions in the classroom is a major deficiency in the student-based framework outcomes. Educators should also take care to read the descriptions of the framework's elements supplied within the supporting article because the intended purpose of each step is not always clear from the framework image alone.

If educators are not looking for a conceptual framework or a framework for improving pedagogy, then they may prefer a framework or tool for their students to use. If this is the case, they might like to consider the Beghetto framework (2002). This framework outlines using CT to critically consider assumptions about creativity. Educators may not initially perceive the relevance of this tool if they are not interested in creativity, however as later explained it can be adapted to using CT to explore other constructs. The framework has four phases designed to help students: recognise existing and perhaps "static" views; confront "alternative understandings"; evaluate and understanding the benefits and limitations of their perspective in the context of their future profession; and generate a stance on creativity based on this understanding and evidence (Beghetto, 2002, p.35). Even though creativity is a separate construct to CT, it has been included here because of its relevance to the Australian Curriculum which groups these two ideas together as part of the general capabilities (ACARA, 2011). Educators wanting to use this framework should note that it was developed for a higher education setting to give students the opportunity to reflect on their "implicit understanding of creativity" and their intended profession which Beghetto notes may have "never been examined" (2002, p.35). However, the phases in the framework could easily be adapted to consider constructs other than creativity, so long as the rest of the details about the process remain the same. For example, students could use the four phases to examine how CT (or any alternative construct) applies to their future professions; how creativity applies to their studies (or alternative contexts); or to further generalise the overall journey - how construct X applies in context Y. Overall this framework contributes an example of applying the process of CT to pre-existing beliefs with the purpose of challenging the beliefs and refining them.

Lastly, a framework that educators may be unaware of is a two-phase framework Edwards created to help promote CT in nursing (Edwards, 2007). It has been included in this review because nursing education has a long-standing inclusion of CT in their classrooms since nurses need to “deal with rapid change” (p.303) and generate and justify solutions for situations with complex criteria on a daily basis. However, the real value of this framework to non-nursing educators is that it is thorough and actionable because the model was generated to be rigorous enough support sound decision making in the life and death contexts that nurses encounter in their workplace. The framework has 12 stages that are broken into a decision-making phase, and a reflection phase. The framework is self-described as “flexible and dynamic” (p.309). It highlights the cyclical nature of CT, and how the judgements and generation of knowledge through one cycle can strongly influence the opportunities in the next cycle. For example, Edwards notes that the evaluation of phase 1’s solution in phase 2 “may encourage [the generation of] new policies and procedures” (p.309). This generation of new ideas can then affect phase 1, by adding to the list of possible alternatives for nurses to choose from when determining the best treatment plan to follow. Phase 2 also involves nurses evaluating the outcomes of their previous decisions and actions, which serves to increase that individual’s subjective and objective nursing knowledge for the next round and may help them to recognise gaps in their previous decision-making process. It also helps them to improve their thinking. Overall this framework is useful, because it models both the judgment process and the reflection process of CT, whereas other approaches or definitions may only emphasise one of these. The summary used to outline the framework (see Edwards, 2007, p.308) is also advantageous because the language is accessible and only needs minor adaptations to be made applicable to areas other than nursing. For example, by replacing ‘nursing knowledge’ in phase 1: step 3 with a more relevant discipline area. Phase 1: step 6 (concerning potential sources of conflicting values and approaches) would also need a few adjustments to direct students’ thinking towards ideas that are appropriate to the adapted context. The main criticism of the Edwards (2007) framework is that it does not promote itself as adaptable, so non-nursing educators may overlook it. Especially since perception surveys such as Shell (2001) and Stedman and Adams (2012) have indicated that educators have a limited amount of time to search for information on CT and are often untrained in CT development (so they may not know how to analyse frameworks or definitions especially those that do not invite them to do so). The perception surveys also indicated that educators gather working knowledge along the way (Black, 2009), so this means they need conceptions that are easy to understand and tools and frameworks that are easy to use.

Frameworks are effective when actualising, disseminating and transforming concepts and tropes such as CT. They are capable of containing lots of different aspects related to CT. They can elucidate the components, the process, and the expectations to determine if CT has occurred. Further, because frameworks are actionable, they can help overcome the barrier caused by the disconnection between theory and practice. In general, the major criticism of all of these frameworks is that many are based on a particular definition of CT. This is problematic because there is still not a consensus definition on CT, which means these frameworks may not be widely applicable, especially if the educator has a different appreciation of the construct of CT. In some cases, these frameworks also introduce confusion about CT by adding extra layers to CT rather than simplifying it; or they oversimplify it, failing to justify why it excludes an aspect that other theorists think is important. Until these frameworks and definitions are capable of dealing with the areas of contention about CT, doubt about CT's definition will remain. So even with the extensive literature on CT, because of unsettled opinions, there remains a need for some additional purposeful, reasoned thinking on CT (specifically with regard to how to navigate, organise and explain the CT process generally and also explain how it might vary in different disciplines). The new definition and framework introduced in the next Section 1.2.4 (but discussed extensively in Chapter 2) was developed to address these needs by supplying terminology and organisation to the construct while leaving the details open to adaption by the educator.

1.2.4 Product of CT, process of CT and a new CT framework

Despite multifarious and sometimes polarising views about CT, recent reviews have improved the degree of clarity (Bailin, 2002; Moon, 2007; Lai, 2011a; Niu et al., 2013). Bailin (2002) provides a science perspective and makes a strong argument for the fact that it is the "quality" of thought "which distinguishes critical from uncritical thinking" (p.364). Moon (2007) highlights the need for "clarity and precision in language and ideas" and "persistence" noting that – CT is "more than a set of skills and processes" (p.54) and "many adept thinkers can manage without ... learning all the critical thinking skills… and [acquiring] the concepts" (p.54). Lai's review (2011a) gives critical insight into the shared ideas scholars have about CT, as well as highlighting arguments around the general and domain-specific aspects of CT. Niu et al. (2013) provide a definition that merges skills and dispositions, expressing CT as "intellectually engaged, skilful and responsible thinking" that "requires the application of assumptions, knowledge, competence, and the ability to challenge one's own thinking" (Niu et al., 2013, p.115). These four approaches merge key components of the existing philosophical, psychological and educational perspectives on CT.

What is noteworthy about this set of perspectives is that all authors (Bailin, 2002; Moon, 2007; Lai, 2011a; Niu et al., 2013) put a high degree of focus on the 'product' of CT - this being the 'response' that shows evidence of engaging in the process of CT. By adding this 'product focus' into the mix of the assortment of previous definitions it becomes possible to see how it could be useful to approach defining CT as a process, whose stages can be inferred from the product to which it gives rise. But how does this product or response show evidence of engaging in the process of CT? The Paulian approach would consider this thought product to display "clarity, accuracy, precision, relevance, depth, breadth, logical, significance and fairness" (Paul and Elder, 2006a, p.5). However as explained in the previous section, the Paulian definition emphasises standards of thinking, with its purpose for critically thinking being to improve the thinking. While the Paul-Elder framework is a useful tool to judge the quality of the thinking, it fails to fully resolve the non-metacognitive aspects of the product of the CT process. Therefore, to look at the cognitive component (or the act of knowing), consider Facione's perceptive on CT, which has more elements that could be evaluated. This perspective considers CT as a process of "purposeful, self-regulatory judgment" that uses thinking skills to explore the "considerations on which that judgment is based" (1990b, p.3). Through this approach, evidence of engaging in the process would include judgement and explanation. There is also the potential to explore the thinking skills used along the way. By combining Facione's aspects, the notion that judgment without reasoning is opinion; and Peirce's ideas about doubt driving inquiry as it searches to "settle an opinion" (Peirce, 1872, p.6) - it becomes possible to derive CT as reasoned thinking whose purpose is seeking resolution for doubt. If this is the case, then the product of CT is understanding that can be used to resolve uncertainty (or usable knowledge). Further, to be capable of settling an opinion, this usable knowledge would need to be well founded. This would mean exploring the construction and validity of the information as understanding and judgments around the issue are generated, or simply put, being able to explain why you know what you know.

Returning to the idea of describing CT in process form, it becomes possible to see how constructing a framework definition incorporating the purpose, process and product of CT could lead to a clearer explanation than trying to assemble lots of detailed attributes into defining sentences. Using these particular elements also embodies the etymological roots of CT. For example, the Greek word for 'critic' is *kritikos,* meaning being able to make judgements (Harper, 2018a); its Proto-Indo-European roots (krei*) means "to sieve…discriminate, distinguish" (Harper, 2018b); and its adjective form (*krinein*) means "to separate, decide" and "judge" (Harper, 2018a, 2018b). The history of this concept is grounded in having a purpose (sieve and judge), a process (separate, discriminate, distinguish and decide) and a product (judgment). As such, a framework definition has been generated, whereby the purpose is to resolve doubt though inquiry, the process uses thinking skills and criteria to inquire, and the product is a reasoned, reflective judgement pertaining to the inquiry. Or perhaps more eloquently: *Critical thinking is a purposeful inquiry process that involves the deployment of thinking skills and criteria (appropriate to the contex*t) *to make reasoned, reflective judgments and transform information into useable knowledge to resolve points of uncertainty.*

Since the purpose of the literature review is to discuss other scholar's ideas to frame the doctoral research, further explanation around this framework definition and the framework elements are covered in Chapter 2. Instead, the discussion now shifts to the development of CT and the Delphi panel's recommendations (Facione, 1990a, 1990b) about the actionable qualities of CT. However, it is first worth reflecting on the role of this new framework compared to existing constructs.

Flores, Matking, Burbach, Quinn and Harding (2012) suggest that "thinking critically about critical thinking should allow one to process the dialectic nature of various constructs into a more integrative whole" (p.216). My new definition and framework achieve this by helping to unify and incorporate existing ideas about CT rather than adding yet another perspective. They provide a way of organising and incorporating existing ideas into a process form that can be used to reach a clarified view that is **adaptive to context**. This enables the new framework definition to avoid some of the drawbacks of other approaches. The new definition is specific enough to clarify what CT is and why the process is undertaken, yet general enough to be adaptable to pre-existing concepts about each of the elements. So instead of trying to re-invent the wheel, it is intended to enable users to take the existing wheels and construct the car. The non-specificity of the broader characteristics in the definition means they can ensure the car's features suit the user's needs, yet the inclusion of core components means that the car will always remain recognisable as a mode of transport (or in the case of CT, recognisable for its purpose as a mode of thinking). So the power of this new approach is that the contextually specific details do not matter, so long as the broader characteristics (having criteria, being reflective, deploying thinking skills) and the thought product of the judgment process [the deeper thinking, reflection and understanding arising out of the initial thinking or reflection], remain central to determining what constitutes critical and uncritical thought.

Having an adaptable framework of the core and common characteristics of CT **empowers practitioners to choose** the details about the elements that are **relevant to their context**. For example, allowing the discipline area to guide the features about the expected standards of thinking and the qualities of valid reasoning and judgement increases the applicability of the CT process to the learners as well as improving the quality and validity of the resulting judgements for that context. This approach also has the advantage of generating a more contextually appropriate and valid thought product, meaning a stronger discipline-based understanding about the issue under inquiry will be generated. Given both the variation in demand for particular thinking skills in different learning domains and the multifaceted nature of CT, this new framework may be as close to a consensus on the construct of CT as can be achieved.

1.3 Developing CT skills

The ongoing discourse about the nature of CT extends into the development of CT skills and dispositions (Kuhn, 2019). Theorists have contemplated both the context and methods for CT development (Norris, 1989; Lai, 2011a). There has been ongoing discussion about the influence of implicit versus explicit teaching approaches (Sternberg, 1986 and 1987; Ennis, 1989; Halpern, 1999; Abrami et al., 2008; Marin and Halpern, 2010; Abrami et al., 2015); and if educators should focus on the skills generally (Halpern, 2001; Van Gelder, 2005; Robinson, 2011); the skills integrated with discipline-specific content (McPeck, 1981; Paul, 1992; Willingham, 2007); or using a “mixed-model” approach (Sternberg, 1987, p.255). However, with varying levels of success from similar development approaches (Abrami et al., 2008, 2015), and conflicting data (Gellin, 2003; Behar-Horenstein and Niu, 2011), the development of CT seems elusive. This section considers the nature of CT development; including the different theoretical perspectives as well as barriers to developing CT; various approaches for developing CT; and CT development in Australian education.

1.3.1 Is CT developable?

CT is seen as a skill-set that is important for the success of future generations (Pithers and Soden, 2000; Shehab and Nussbaum, 2015; Sellars et al., 2018), therefore there is a general acceptance that it can and should be developed in all individuals through their educational experiences (Dewey, 1910; Facione, 1990a; Ennis, 2018). However, some of the philosophical rhetoric about the qualities of a critical thinker portray CT as more of a fixed or innate ability, expecting consistent deployment and displays of CT by critical thinkers for them to be named so (Facione, 1990b; Paul, 1992; Paul, 2005). In addition, some theorists such as Orstein and Hunkins (2004) have suggested CT development can only occur at the formal operational Piagetian stage since it requires abstract reasoning, and therefore if the student has not achieved that stage they will not be able to think critically. However, Shillady (2011) believes that "children are investigators – born with an innate desire to explore and understand the world" and display dispositions often associated with CT such as "curiosity" and "inquisitiveness" (2011, p.12). Further, in a study by Bascandziev and Harris (2010) 3.5-year-old children were found to improve the accuracy of their predictions through visual and verbal training in cause and effect relationships, so not only are pre-operational stage children both fascinated by and capable of problem-solving, but they can also improve at it. While Piagetian perspectives such as Orstein and Hunkins' (2004) could imply that CT is not something that can be developed, this belief is not held by the majority of theorists. In fact, "teaching thinking" has been a "long-term aspiration" in education (Glaser, 1983, p.30). When examining these arguments about what it means to be a 'critical thinker' it is important to note that the dialogue is actually focused on the deployment of CT skills across time and space, not how they might be acquired. Whereas discussion about the development of a critical thinker in the literature tends to focus on 'what, where and how' the various components of CT can be developed. The dialogue is more about how to enhance the development of these skills and dispositions, not if they are developable at all.

1.3.2 Developmental considerations regarding the nature of CT

There are common ideas about the core thinking skills educators desire students to possess (see Definitions Section 1.2). However, there are different ideas about the ways to help students acquire and develop them. The main approaches concerning CT development were first classified by Ennis (1989) and are often referred to as Ennis's typology of instruction (Behar-Horestein and Niu, 2011). These classifications include a general approach, two variations of an embedded approach, and a mixed approach. The general approach overtly and only teaches for skills and dispositions related to CT (Ennis, 1989, p.4). This can be done in concrete or abstract form. The concrete format presents general CT principles explicitly and explores their application to content (which could be subject-specific or just general examples); whereas the abstract format just focuses on general CT principles without exploring their application to content. The embedded approaches, implant skills and dispositions within course content, but vary by whether the general principles are explicit or not (Ennis, 1989, p.5). Lastly, the mixed approach blends the explicit teaching of CT skills and dispositions with subject-specific school content and "non-school" contexts (p.4). For a summary of the core differences between these instructional approaches see Table 1.4.

Table 1.4 Summary of the differences between the instructional approaches in Ennis's typology. *Note.* Reprinted from 'Critical thinking and subject specificity: Clarification and needed research' (Ennis, 1989, p.5)

Removed due to copyright restriction

Recent evidence shows that to maximise the chance of successful deployment and execution of CT skills, a combination of general and discipline-specific learning opportunities need to be provided (Abrami et al., 2015). This would be consistent with a mixed approach according to Ennis's typology. These findings complement the Delphi panel's recommendation that educators should "be guided by a holistic conceptualisation" (Facione, 1990a, p.4) to create learners who can successfully integrate and execute CT skills in their "studies" and "everyday lives" (p.4). However, because Abrami et al.'s (2015) meta-analysis results are recent (relative to the publication cycle), the impact of this summative evidence is not yet seen in the approaches taken in the literature. Instead, existing classroom-based studies are formulated from a vast number of definitions and approaches, with some focusing on general skills or dispositions; some using embedding CT attributes within content; and others using a more blended style. However, there is another matter to explore which receives a lot of attention in the literature regarding choices around instructional approaches. It can help unpack why the mixed approach produced the greatest effect in the Abrami et al. meta-analyses (2008, 2015). It pertains to the role of general and domain-specific content and the deployment and transfer of CT skills.

The facets of CT include knowledge, thinking and reasoning (Glaser, 1983, 1984). There is no way around acknowledging that it involves a number of different cognitive skills which are valued in multiple settings. But understanding the development of CT is increasingly complicated by the fact that these CT skills become conjoining cognitive processes to reach a point of discerning judgment. As a result, it becomes unclear about if and how to separate the parts of CT, whilst allowing it to remain 'reasoned thinking', and not just 'thinking' or 'undiscerning judgment.' A large portion of the literature is dedicated to discussing instructional approaches and whether CT is something that should be taught exclusively from subject matter; along with subject matter; or embedded in subject matter. However, what this debate really revolves around is the relevance and application of CT skills across learning domains. Generally, the arguments get broken down into general and discipline-specific ideas concerning the development and transfer of CT skills. Is it a general ability as Halpern believes (2001); is it domain-specific ability (McPeck, 1981); or is it a general ability that is more readily developable in domain-specific contexts?

Teaching for CT as a general ability

Generalists believe that CT consists of a set of generic skills, abilities and disciplines which can be applied across a broad range of contexts and circumstances (Bailin and Siegel, 2003). Generalists such as Paul (1985), Halpern (2001), and Van Gelder (2005) contend that students need deliberate practice in exercising CT skills and abilities, and imply that this type of practice can be best achieved when CT is taught as a separate and explicit part of the curriculum. In a position paper, Greiff et al. (2014) suggested education in the twenty-first century "should equip students with domain-general problem-solving skills in addition to domain-specific factual knowledge and problem solving strategies" (p.74). However, Greiff et al. (2014) argued that "contemporary education systems fall short" (p.74) in equipping students with domain-general aspects because the "discipline boundaries" impede the amount of attention given to "cross-curricular skills … general conceptual frameworks and intellectual skills [such as information processing, reasoning, self-regulation, metastrategic thinking, decision making etc.]" (p.75). Greiff et al. (2014) does not use the term CT in their paper; however, the descriptors and cognitive abilities they assign to "domain-general problem solving skills" are comparable if not the same as the components of CT as described by Facione (1990a, 1990b) and other theorists (Ennis, 1991; Paul and Elder, 2006a).

In the generalist perspective, CT can be considered as an "umbrella term," which in its normative sense refers to how thinking is carried out (Bailin and Siegel, 2003, p.188). However, one main problem with, and criticism of, this approach is achieving the successful deployment and application of the cross-curricular cognitive abilities from the general context into the domain-specific ones. There is uncertainty about whether this requires intervention [scaffolding or training in domain-specific areas], or if students have the intellectual flexibility to do this themselves? The Organisation for Economic Co-operation and Development (2014) acknowledged that mastery of reasoning skills is evident when students are “motivated to engage with unfamiliar problems” and can “solve problems [outside of their expertise] efficiently” (Organisation for Economic Co-operation and Development [OECD], 2014, p.29). However, as previously discussed by a number of theorists (Brown, Bransford,, Ferrara and Campione, 1982; Glaser, 1983, 1984; Ennis, 1989; McPeck, 1990; Halpern, 1998) getting this effective and efficient transfer of CT abilities to occur can be challenging. More recently, Willingham (2007) noted that students may exhibit CT skills and abilities in one domain, but fail to do so in another. Angeli and Valanides (2009) found that general approaches have low transfer success unless the particular context had been scaffolded for the student because the students compartmentalise the knowledge they have gained to that general context. So it seems the generalist approach leaves itself open to the possibility of poor transfer of CT skills unless some training also occurs in domain-specific areas.

Lastly, this generalist view is criticised because what constitutes valid evidence and knowledge varies with the context (Bailin, 2002; Weinstock et al., 2017). Dunne (2015) posits that it is difficult to explore how generalisable CT is because there is a need to consider both relative and absolute positions. He questioned whether or not the application of the same criteria by two individuals to the same ethical dilemma would lead to the same conclusion. Furthermore, he reflected on whether "the merits of reasons are the same in all contexts" (2015, p.90) and if time plays a factor. Discipline or domain-specific criteria could bring clarity to Dunne's ponderings because they bring in rules about how general thinking criteria should be applied and show how reasoning processes can be different but still valid. Ennis aptly notes that "CT cannot occur without "content" but you can "teach content-free" CT principles (1989, p.9), recognising that it is the context, not the content which frames the intellectual demands and achievement standards in that learning domain. In education settings, the ‘context' is readily framed through subject-specific content and criteria. This role of subject-specific awareness in CT is explained more in the next section.

Teaching for CT as a domain or subject specific skill

When reporting the Delphi panel’s findings, Facione commented that “too much of value is lost if critical thinking is seen as a list of logical operations and if domain-specific knowledge is seen as an aggregation of information” (1990a, p.5). Domain-specifists’ note that the essence of CT is “rationality” (Siegel, 1988), and to employ “good judgment” (Lipman, 1987, p.39), and identify that knowledge and context are crucial to this discerning process. Proponents of this domain-specific approach believe there needs to be a content-based or 'discipline-specific' aspect to CT. This helps to set the frame for contextually relevant knowledge and skills (Ennis, 1989; McPeck, 1990; Weinstein, 1995; Willingham, 2008; Weinstein et al., 2017). Domain-specifists place a high value on the niche variation in the criteria and standards across disciplines and how they govern what constitutes 'sound decision making and judgment' within different fields.

Green and Yu (2016) have reflected that "successfully addressing" certain concepts (understanding the personal and global impacts of political policies when deciding who to vote for), can require "deep disciplinary knowledge" and also awareness of "how experts in those disciplines engage" within the content of their field (p.46). Ku and Ho (2010) found epistemic knowledge to be important for metacognitive strategies noting that "having only an awareness for the need to apply metacognitive strategy is not enough for good performance; one must also know when, how and which strategy to use at different contexts" (p.253). The same is true of CT – if general awareness of CT strategies was sufficient to facilitate the appropriate deployment of the CT strategies within any learning domain, there would likely be more evidence of success from general CT courses. Yet the evidence does not reflect this transfer (Angeli and Valanides, 2009). However, it is really not surprising that transfer from general courses is low, because the constraints that shift 'thinking' into 'critical' thinking in a specific learning domain come from specific contextual knowledge, so it is almost unreasonable to expect that a student trained in the generic skills could transfer these abilities without also having awareness of the applications and constraints in the more specific domain (Brown et al., 1982; Dumitru, 2012; Wall, 2015).

Ultimately, domain-specifists believe that each field maintains a unique set of parameters that define high-quality thinking (Green and Yu, 2016), so even though the core thinking traits deployed in the process of decision making and judgments might be the same as other learning domains, domain-specifists would argue that it is the specific criteria that shapes and equates to a sound thinking in that field which demands some degree of specific knowledge to produce the most ideal outcome (Possin, 2008; Robinson, 2011). Golding (2011) went so far as to try to "side-step the [general ability/ discipline-specific] debate" (p.360) when generating a skills-based approach to CT development, yet the resulting pedagogical strategy he put forward was "based on discipline-specific" thinking (p.361). Golding does make note of this but also commented that his method "applies equally well" to both approaches (p.360). However, because he did not supply a definition nor a means of assessment for his approach, the question of 'how can one compare the success of his method in either a domain-general or discipline-specific setting?' remains unresolved, and so the argument for discipline-specific learning remains. Yet the majority of this rhetoric on domain-general courses versus domain-specific teaching fails to recognise that it may be a false dichotomy. Davies (2006) argued that generalist and specific approaches are complementary, and further suggests that making it a two-sided choice involves a "fallacy of false alternatives" (p.180). Greiff et al. (2014) acknowledged the role of both domain-specific and domain-general problem- solving skills, but expressed that more effort should be directed toward developing "cross-curricular" (generalisable) skills (p.80). Ennis, a generalist, has also reflected on this “either-or” position and now represents these concepts on a “rough continuum with clear examples at either end” (2016, p.30) noting that there is a role for both sets of abilities and dispositions. The contributions each perspective could offer are explored further below.

Blended approach to CT development

The struggle to generate a consensus definition of CT, and the ongoing discussion about the domain-general and domain-specific attributes of CT, highlight that CT is a process with both broad and specific applications. There is value in possessing a range of CT skills and knowing where they can be applied, just as it is helpful to possess other generic capabilities such as how to read, write and count (Facione, 1990a). As Sternberg (1986) pointed out, the language may vary but most authors would agree that there is a common set of thinking traits which assist in sound decision making and judgments no matter the discipline. It is in the finer details where the disagreements exist, specifically in the role of a learning domain in shaping the values, criteria and standards that amount to ‘sound judgment.’

Glaser (1983, 1984), McPeck (1990) and Facione (1990b) have acknowledged that whilst CT skills are general enough to transcend disciplines/subjects, domain-specific knowledge is sometimes required to successfully deploy CT skills in specific contexts and make reasonable judgments. Puolimatka (2003) further notes the value of discipline-specific knowledge commenting that “one cannot be critical in any field without being closely acquainted with it” because “one cannot employ skills of critical thinking constructively and creatively” without considering “existing knowledge” (p.11). This statement is not tying any specific CT skills to a discipline (the skills themselves are general) however it expresses that the effectiveness of their application is related to knowing which ideas to challenge. That is, knowledge and knowledge of one’s ignorance is what CT acts on and from, so some awareness of that disciplines foundational knowledge is needed to undertake and produce thinking which is acceptable to that field’s criteria and standards. Glaser (1984) recommends “teaching specific knowledge domains in interactive, interrogative ways so that general self-regulatory skills are exercised in the course of acquiring domain-related knowledge” (p.30). This perspective recognises that the crossover of the thinking skill set presents an opportunity to scaffold generic skills whilst gaining discipline-specific knowledge. To this end, domain-general CT courses could offer a platform for discipline-specific areas to work from; or CT could be taught both generally and through domain-specific examples within one discipline-based course. However, Golding’s approach (2011) sat somewhere between these two.

Golding (2011) focused on educating for thinking skills such as "inquiring, problem-solving, argument analysis and construction, uncovering and evaluating assumptions, justification, interpretation, and questioning received wisdom" (2011, p.360) to develop what Perkins' would term ‘reflective intelligence' (Perkins, 1995). Golding presented a modified version of Paul's (1995) Socratic questioning strategy as the vehicle for developing CT skills, which he feels "creates an educative community of CT where students ask and answer thought encouraging questions" (p.361). Golding's Socratic questioning method is different. Unlike Paul's method, they are not intended to lead students to "understand predetermined content" (p.365), instead, it uses a community of inquiry to educate for CT skills, dispositions, epistemic understanding, content knowledge and the criteria for successful CT (Golding, 2011). Golding's method reframes the value of content – highlighting that "know how" (skills) or "knowing about" (content) are not sufficient to make students a critical thinker (p.358). However, these notions of open-ended content do not suit the ‘breath of knowledge approach' displayed through most education systems. As a result, Golding's methods present content challenges many educators may not wish to face at this time. Further, Golding's article is entirely theoretical, and this is another concern, as educators want to know how theory translates to the classroom, in addition to how it can be assessed and what results are achieved.

Returning to broader developmental considerations, Sternberg (1985), Ennis (1997), Facione (1990a, 1990b), Paul (1992), and Kennedy et al. (1991) are all advocates of a mixed approach, and there is growing evidence to show this practice produces the largest effect on CT development (Abrami et al., 2008, 2015). Facione noted that “the experts could not deny one of the best ways to learn CT is within a subject context” (p.14), however, the experts also suggested that because CT has “application in all areas of life” (p.4) “a solid liberal education” (p.5) using “explicit” and “direct instruction” (p.14) of CT in different contexts would provide the best opportunities for students to develop into critical thinkers. Kennedy et al. (1991) recommended using a mixed approach after reviewing existing research and concluding that there is not sufficient evidence to support the superiority of any one method. In their meta-analyses, Abrami et al. (2008, 2015) identified that a mixed approach (teaching general CT skills with concurrent teaching and scaffolding in discipline-specific contexts) had the largest positive effect on CT development. The literature on transfer also supports the inclusion of domain-specific scaffolding alongside domain-general teaching, as the spontaneous transfer of CT abilities is unlikely (Pithers and Soden, 2000; Willingham, 2007). However, when Abrami et al. (2008) and Behar-Horenstein and Nui (2011) conducted their analyses, they identified many confounding variables which made it complicated to identify the true effectiveness of interventions. Were the effects that were seen due to the learning environment, instructor training, instructor experience, instructor knowledge, student-instructor interactions, student-student interactions, duration of intervention, the assessment tool used or the research design (true experimental, quasi-experimental or pre-experimental)?

Some answers were gained from Abrami's (2015) recent meta-analysis where the group reduced the number of confounding variables to explore the effects of the teaching approaches at a deeper level. In this meta-analysis, a general approach on its own (explicit teaching of CT skills separately from other content) was also found to be mostly ineffective, and an implicit across curriculum approach had the smallest effect size, providing further support for a blended approach. The Delphi panel did not have this evidence base, but through their collective experience were able to foresee that limiting both the breadth of experiences and inclusion of particular CT components would be to the detriment of CT development (see Facione, 1990b, p.10). The panel cautioned that "the education of good thinkers is more than training students to execute a set of cognitive skills" in addition to commenting that an 'either-or' approach to domain-general and domain-specific abilities "truncate its utility... and value" (Facione, 1990a, p.14). However, the evidence to support these recommendations was not found early enough to assist the permeation of these messages and prevent the apparent attrition of CT development in education (Pascarella and Terenzini, 2005). But in light of the accruing evidence supporting the Delphi panel's ideas about instructional approaches, it seems timely to consider whether anything else can be gleaned from the other recommendations the panel made about CT development.

1.3.3 Influence of the Delphi panel's recommendations (Facione, 1990a) and corresponding barriers influencing CT development

The initial purpose of the Delphi panel (Facione, 1990a, 1990b) was to contribute ideas about the development and assessment of CT skills. So while a definition emerged from the various iterations and debate of what is and is not CT, the intended purpose of the study was to guide goal setting to improve CT and inform high-quality education practice. Recommendations (R) according to Facione (1990a), CT instruction should:

- equip students for good thinking and judgment in their lives as well as the classroom (see p.4 – R1).
- provide opportunities for learning the procedures and also when to apply them in different contexts (see p.5 - R3, p.17 – R14).
- model and nurture the critical spirit, not just the cognitive skills (see p.11 – R4, p.13 R5, p.14 – R6, p.14 - R7).
- be an explicit instructional goal throughout the K-12 curriculum (see p.15 – R9, p.15 - R10).
- have minimum proficiency expectations to proceed through each education level and stage (see p.16 - R11).
- foster confidence in students' own powers of reason, rather than dependency on rote learning (see p.18 - R15).

Many classrooms are characterised by rote learning and assessment, and focus on cognitive skills and accepting knowledge rather than challenging it (Paul, 2005). Commentators have noticed that "teachers are having difficulty teaching for CT" (Shell, 2001, p.287), and "students may be more interested in the grading practices of their teachers" rather than gaining the "skills necessary for effective learning" (Choy and Cheah, 2009, p.199). There is a perception that the deliberate development of thinking skills is "rare" (Vainikainen et al., 2015, p.54). This apparent lack of penetration of Facione's recommendations into classrooms is disappointing, especially since it contrasts to the widespread use of the Delphi-panel's consensus definition. In many ways, it is similar to Bloom's educational handbook being "one of the most widely cited yet least read books in American Education" (Anderson, Sosniak, and Bloom, 1994, p.9). However, Paul (2005) has described three "serious" obstacles impacting the successful development of CT and inhibiting "long-term institutional change" (p.27) – educator ignorance of a "substantive concept of" CT (p.27); educator ignorance of the deficit in their understanding of CT and subsequent teaching for CT; and ongoing use of traditional teaching methods, despite "reform efforts" (p.27). What is keeping these obstacles in place?

Williams and Burden (1997) explain that an educator's perception of their role in the classroom – as either facilitators of learning or disseminators of information – shapes how they incorporate CT into the classroom. The fact that perceived content coverage demands are frequently cited as a barrier to including more CT development time in the classroom, indicates that the majority of educators see themselves (or their priorities) as disseminators of information. However, what this perspective fails to note is that this classroom-identity is not necessarily self-generated. It may be an artefact of the education system because if CT development is not valued at an institutional level it makes it harder for educators to prioritise it. For example, Aliakbari and Sadeghdaghighi (2012) identified that 60-78% Iranian educators surveyed felt that CT was not a "primary objective of their teaching," nor a "university priority" (p.4). Prioritisation of 'information dissemination' rather than an identity of 'learning facilitation' could be responsible for the lack of proficiency in CT. However, failure to overcome these obstacles and successfully incorporate intentional CT development into education could also be due to the majority of educators not understanding how their teaching methods may be impacting CT in the first place. Yet these educators may not realise that by not providing sufficient space for thinking about content, that very content is either "unlearned or mis-learned" (Paul, 2011, p.19). They may believe that "whatever problem exists in the learning process is the fault of the students or beyond their control" (Paul, 2011, p.34).

Reflecting on his experiences in education, Paul (2005) expressed that "few faculty recognise what it takes to transform instruction so that students use their thinking to take ownership of course content" or to "think analytically" about content (p.36). For example, in a study of educators in California, it was found that only 9% "were clearly teaching for critical thinking on a typical day" despite 89% claiming CT as "a primary objective of their instruction" (Paul et al., 1997a, p.18). Schneider and Miller (2005) reported similar findings with 93% of academics indicating CT was a course focus, but only 6% of seniors were proficient at CT (as identified through standardised testing). Aliakbari and Sadeghdaghighi (2012) identified only 10% could "explain their departments definition of" CT (p.3). Black (2009) identified that many UK educators teaching CT courses have no prior specialist CT subject knowledge, instead acquire working knowledge along the way. Other perception studies (Shell, 2001; Choy and Cheah, 2009; Stedman and Adams, 2012; Alwadai, 2014) found that faculty members lacked knowledge and understanding about CT. Previously Paul (2005) suggested "when faculty have a vague notion of [CT] they are largely unable to identify ineffective teaching practices or develop more effective ones" (p.27). This was confirmed by Abrami et al. (2015) who identified that teacher training in CT methods was an important factor in successful CT development, and those without training were generally unsuccessful. So it is likely this lack of awareness and understanding could be responsible for maintaining a culture of recall rather than reasoned thinking. But why are teachers unaware of CT and its methods, given that there is so much scholarship on CT?

Despite the efforts of the Delphi panel, the ongoing lack of consensus definition (see Section 1.1) could be compounding the general ignorance. Educators might be unsure how to navigate through the various theories and opinions in the literature; or might be too time poor to keep up and modify their practice in line with more recent findings (such as those by Abrami et al., 2015). A number of researchers have investigated educator perceptions of CT to identify why development continues to remain a challenging task for educators (Paul et al., 1997a, 1997b; Shell, 2001; Schneider and Miller, 2005; Choy and Cheah, 2009; Black, 2009; Aliakbari and Sadeghdaghighi, 2012; Kowalczyk et al., 2012; Stedman and Adams, 2012; Aldwadai, 2014). With the main barriers identified as time (for both developing CT content, and also teaching CT content); student resistance (motivation, and engagement with CT teaching approaches); training (limited knowledge or low confidence about CT and effective teaching strategies); plus, a lack of resources or support. These barriers seem to be similar across education systems, education levels, and teaching domains. For example, Black (2009) reported that UK school teachers faced a number of barriers when introducing CT into schools including: working from insufficient training and resources; minimal timetable allocation compared to other courses they teach making CT difficult to prioritise; needing to change student motivation (and perceptions of the difficulty about the course). Similarly, Kowalczyk et al. (2012) identified access to appropriate teaching materials; high workload; a lack of student motivation as well as resistance to CT teaching methods; and overall confidence in skill levels for implementing the CT strategies, as major obstacles when investigating barriers to implementing CT development strategies among nursing program directors in the US. Paul et al. (1997a, 1997b) found that there was a need to increase educator awareness of the conceptions of CT; awareness of methods for incorporating CT into the classroom (both general, and within discipline areas); and also improving support and training in CT (for current and prospective educators), when they surveyed Education, Art and Science faculty at universities across California. So how is this lack of awareness and support, in addition to other barriers, impacting educational outcomes?

Pascarella and Terenzini investigated CT gains from college (1991, 2005) and found that college does improve CT, however, the magnitude of the effect was smaller in the more recent study. Combining Pascarella and Terenzini's results with their own investigations of college outcomes, Arum and Roksa felt this might indicate that higher education is becoming less effective for developing CT (Arum and Roksa, 2011; Arum and Roksa, 2014). This does not reflect well on the dissemination of Facione's work with the Delphi panel (1990a, 1990b), nor the perspectives that have subsequently emerged (Ennis, 1991; Lipman, 1995; Paul, 1996; Facione, 2000; Hatcher, 2000; Paul and Elder, 2006a; Beghetto, 2002; Possin, 2002; Lipman, 2003; Paul, 2005; Duron et al., 2006; Edwards, 2007; Almedia and Franco, 2011; Ennis, 2011a; Paul, 2011, 2012; Rabu et al., 2013; Dwyer et al., 2014; Facione, 2015) – since instead of seeing an improvement in CT development, there has actually been a decline. However, this apparent attrition may be a consequence of a few things.

Firstly, Pascarella and Terenzini's (1991, 2005) method changed. They included a broader range of assessment tools in their second study (2005). This makes their two studies less comparable because Hatcher (2013) found that different CT measures can result in different effect sizes even when the course has remained unchanged. The likelihood of this is increased because Pascarella and Terenzini's calculations involved comparing standard deviations, so the potential for increased variability around the mean would have changed with the inclusion of different instruments.

Secondly, it could be due to language sensitivity or perhaps a language shift, rather than a reduction of CT in the curriculum. Perception studies have indicated that many educators are not trained in the methods of CT, so it is likely that they may not be using the language of CT nor understand why CT is not an emergent property of education. A consequence of this could be what Mayer and Wittrock (2006) have described as the "hidden curriculum" (p.296) in which educators expect problem-solving, but do not explain or teach for it. In these instances, the transfer of CT skills onto generalised assessments is likely to go astray - students may misinterpret questions because they do not know the language of thinking; they may be unaware they have the tools to solve the problems; or they may think they have appropriately and sufficiently responded to the question, but have actually done so in an uncritical way. For example, Manalo and Sheppard (2016) found that proficiency in English had implications for CT performance for university-level second-language learners when implementing CT strategies in the learner's second language. They went on to suggest that "students need instruction on the specific language forms and structures to use to demonstrate critical thinking in their written work" (p.41), which has broader implications in light of the apparent decline of CT language in the enacted curriculum.

But does this issue rest solely with the educators or the education system? What level of ownership do students need to take for their learning, particularly in higher education? From a cognitive perspective, studying CT can pose more of a challenge than studying other content because it is more abstract. The thinking skills required are often characterised as higher-order processes (Bloom et al., 1956; King et al., 1997; Barnett and Francis, 2012), they take time to develop (Wall, 2015) and the prevalence of recall learning strategies will not help students gain these skills. Many of these things are in educator control, so it would not be entirely unreasonable to expect reforms to programming, or on a broader scale, to the education system. However, in a number of studies on educators' perceptions of CT development, educators indicated issues with student engagement/participation and motivation when educators have tried to introduce CT activities into the classroom (see Shell, 2001; Choy and Cheah, 2009; Black, 2009; Aliakbari and Sadeghdaghighi, 2012; Kowalczyk et al., 2012; Alwadai, 2014). There is a student aspect to this as well.

Weinstoc et al. (2017) suggests the disposition to exercise CT should not be underestimated, because its deployment requires wilful intention and purpose. Miele and Wigfield (2014) explored willingness to engage in critical analysis and found that "positive beliefs about their ability to accomplish tasks that require critical-analytic thinking, the extent to which they value these tasks, and the goals they want to achieve by completing them" (p.522) were all important in motivating students. Silvia and Sanders (2010) suggest "interest is central to intrinsic motivation for learning" (p.242); interest also encourages deeper processing (Krapp, 1999). Dweck, Walton and Cohen (2014) explain that "motivational" factors "can matter even more than cognitive factors" when it comes to "academic performance" (p.2). These findings indicate that the dispositional aspects of CT, particularly self-reflection and if necessary self-correction, need to become a part of the general experiences in classrooms. This will help students see the purpose of their thinking and learning and become more interested in it. Whereas many of the current approaches used for developing CT focus on skills, and the dispositional aspects are treated more like existing or an emergent set of traits, rather than something which could be targeted. However, Wall (2015) suggests it is not the skills, but rather the "habit of using them" that will transfer (p.238). Educators need to spend time fostering CT habits, not just CT skills. Dispositions have been a core component to definitions of CT since early conceptions, with the "reflective thought" (1910, p.2) forming the basis of Dewey's perspective on CT.

Numerous authors have highlighted the importance of reflection as a part of CT (Facione, 1990a; Ennis, 1991; Finocchiaro, 1996; Halpern, 1998; Facione, 2000; Paul, 2005). Yet there is still a high degree of uncertainty about the development of these reflective habits, as they sit in the broad field of metacognition, which like CT has a lot of theory and terminology for educators to consider. Some of these ideas are explored below.

1.3.4 Metacognition and CT development

When developing their consensus on CT, Facione's panel members (1990a, 1990b) acknowledged the role of self in CT, noting that self-examination and self-correction are important components of CT. These self-management components are synonymous with aspects of metacognition – another ill-defined concept with over 30 years of theorising (Akturk and Sahin, 2011). Metacognition is often described as thinking about one's thinking (Flavell, 1979), and is a monitoring strategy (Kuhn, 1999; Ku and Ho, 2010) that can also include planning and evaluation among other skills (Brown et al., 1982). It includes knowledge of cognition and regulation of cognition (Schraw and Dennison, 1994); and these metacognitive processes are considered important for optimal thinking performance (Schraw, 1998).

Metacognitive knowledge (knowledge of cognition) is argued to improve thinking performance by allowing problem solvers to better encode and represent the task in a problem context (Davidson and Sternberg, 1998). Whereas, cognitive regulation helps facilitate learning (Schraw and Dennison, 1994) by way of helping the learner to organise, attend to and evaluate their thinking (Borkowski, 1996). Cognitive regulation is also thought to help with the management of affective states (Cross and Paris, 1988; Martinez, 2006).

Metacognitive processes are seen as distinct from cognitive ones because they bring awareness and regulation, but not necessarily task fulfilment (Schraw, 1998; Kuhn, 2000). Metacognition can occur before cognitive activities (planning), after cognitive activities (evaluation), or the processes can coincide (monitoring thinking while completing a task) (Akturk and Sahin, 2011). Kuhn (1999) views metacognitive skills as "second-order" skills that "entail knowing about … knowing" rather than just "knowing" (p.17). However, it is not as clear where metacognition sits in relation to CT because CT has both cognitive and dispositional components. Kuhn (1999) sees metacognition as "central to critical thinking" (p.23), providing a framework for processing and responding to information. Martinez (2006) sees CT as a type of metacognition. Whereas, Schraw, Crippen and Hartley (2006) group CT and metacognition under the umbrella construct of "self-regulated learning," which is the "ability to understand and control our learning environments" (p.111). Pintrich, Marx and Boyle (1993) provided seminal understanding about the link between motivation (self-efficacy and interest) and engagement with metacognitive evaluation. Follmer and Sperling (2016) later confirmed metacognition's mediating role in self-regulated learning. However, the same has not yet been done for CT.

Differences in opinion still exist about the hierarchy of executive functions, depending on whether the metacognitive processes are considered as thought monitors (acting on the thinking), or if the cognitive CT skills are the way to achieve the thought monitoring (more of a metacognitive loop where the CT skills drive more thinking and checking). However, Paul (2005) suggests that CT "moves back and forth between cognitive" and "meta-cognitive" states (p.5). Regardless of whether metacognition acts on, or from CT, essentially CT can be thought of as a process for challenging knowledge. The literature suggests that expert thinkers would be efficient and thorough at doing this; reflecting on existing ideas and evidence, but also contemplating potential weaknesses in one's own thinking when it comes to making final evaluations and judgments. It therefore follows that: CT involves both thinking about knowledge and thinking about thinking. Yet the necessity of self-reflection in order to **do CT** remains to be seen – as the cognitive and dispositional elements are still treated distinctly and reflection is not automatically needed to use cognitive CT skills. However, these self-regulative components are necessary to **be a critical thinker** because this personified version of CT would include the dispositional traits and therefore implies both willingness and expertise to deploy the skills. For example, trade apprenticeships demonstrate how trainees need more than the technical skill to be proficient at that trade – they also learn the "customs and practices", the "hierarchy's" and the "wider context of the labour and marketplace" during their apprenticeship (Murray, 2002, p.1). Educators could learn from this approach (developing student's dispositions to think critically and reflectively, in addition to the product-based skills) to gain insight into how to create critical thinkers who can apply their knowledge and skills beyond the classroom. As Facione (1990a, 1990b) suggested, educators have a responsibility to provide experiences to help students understand the customs and practices so they "can integrate successful execution of… [CT] skills… with the confidence, inclination and good judgments" to employ the tools in "their other studies and everyday lives" (Facione, 1990b, p.4). But achieving this with the self-reflection and metacognitive aspects of CT remains a challenge.

Metacognition is considered teachable or at the very least developable (see Cross and Paris, 1988; Haller, Child and Walberg, 1988; Hennessey, 1999; Kramarski and Mevarech, 2003; Dignath and Büttner 2008). Instruction is aimed at increasing awareness rather than focusing on performance (Kuhn, 2000) and scaffolding explicit strategies (Cross and Paris, 1988; Schraw et al., 2006; Kadian, 2016) with checklists (Schraw, 1998) or questions (Kramarski and Mevarech, 2003). A study by Pressley, Wharton-McDonald, Mistretta-Hampton and Echevarria (1998) found that students who had knowledge of different reflective strategies, as well as specific knowledge about when a strategy should be deployed, were more proficient in their use of metacognitive strategies. Time is also thought to be related to proficiency, with several researchers determining that metacognitive abilities seem to improve with age and exposure (Cross and Paris, 1988; Schraw and Moshman, 1995; Kuhn and Dean, 2004). These findings have a number of implications for CT development.

Firstly, the current, mostly skills-based approach to CT development has a performance-focus rather than awareness-based approach. However, Mueller and Dweck (2008) found that praise for effort in failure was better than praise for intelligence. Dweck et al. (2014) later explored factors affecting academic tenacity and found that it was attributes related to mindsets, goals, self-control and self-regulation which seemed to predict a student's academic achievement, rather than IQ. So perhaps an awareness-based approach, focusing on the dispositions of CT and motivation for thinking, could improve sustained academic performance and CT transfer outcomes because it is more likely to teach students to value and engage in the process of CT.

Secondly, domain-general CT abilities versus domain-specific abilities receive a lot of attention in the literature. To assist with decision making about CT in domain-general and domain-specific contexts, there is a need to understand how performance improves with the deployment of different metacognitive tools. In fact, there is already evidence to show that approaches incorporating both domain-general and domain-specific components produce the greatest effects on CT development (Abrami et al., 2015).

Lastly, the time aspect reinforces that these higher-order functions benefit from continued exposure. Students need multiple opportunities to knowingly engage with CT through their educational experiences - so whilst teaching CT in a specialist course, or even an occasional CT-focused lesson will help build a foundation for CT, they should not form the totality of their experiences with CT.

With its focus on evaluating the standard of one's own thinking, metacognition adds to and improves the CT process. However, CT also has something to add to the construct of metacognition. For example, one main perspective about metacognition is the idea of thinking about one's own thinking typically for monitoring and regulation purposes (Flavell, 1979; Paris and Winograd, 1990). Yet, when it comes to CT and metacognition, limiting thinking about thinking to one's own thinking is contrary to the nature of CT. Metacognitive levels of thoughtfulness could extend to thinking about other's thinking (the external ideas presented to us) - considering whether these external ideas are reasonable and valid, or if it is necessary to challenge them because the creator's thinking contains some bias or fallacy. When describing reflective thinking, Dewey noted that if there is no controversy or doubt about an issue then reflective thinking is not necessary (Dewey, 1929). When teaching students to appropriately challenge knowledge, educators could encourage students to examine the credibility of the source in addition to the arguments, data or evidence provided, to help improve the foundations for the student's resulting judgments.

Like CT, metacognition is multifaceted and theory rich, and this has previously resulted in "fuzziness and generalization" (Tarricone, 2011, p.5). Some of the language and terminology barriers described in Sections 1.2.2 are also problematic in metacognition. For example, Dinsmore, Alexander and Loughlin (2008) identified that researchers were not careful enough about the terminology used in their constructs, or the way they were conceptually and operationally defining metacognition. Yet despite its complexities, Tarricone (2011) has managed to generate a conceptual framework and taxonomy to help with understanding metacognition. The same level of taxonomy has not been yet been constructed for CT, but it demonstrates that this type of approach can be effective for simplifying complex ideas. The next chapter contains a conceptual framework designed to increase the usability of existing perspectives on CT. Whilst this framework is only a step towards the pathway taken by Tarricone, it does provide a starting point for generating a taxonomy - that is a mechanism for categorising existing ideas on CT and provides a way to incorporate and notice CT in education and everyday life.

1.4 Conclusions

Despite numerous efforts by commentators to explain the nature and characteristics of CT and its development, it remains intangible in much of the literature. This review has shown that there is a wide variety of perspectives and a continual search for a unified definition. It revisited the

Delphi panel's instructional recommendations and highlighted how recent literature has supplied an evidence base for some of these conceptions. For example, the Delphi panel recommended providing opportunities for learning CT explicitly coupled with opportunities to apply them in different contexts. Literature published since then presents multiple ways to approach CT development, yet recent findings have shown that a mix of general CT training and discipline-specific conditions lead to the strongest outcomes. The Delphi panel were correct with their beliefs and thinking about this aspect of CT, and this perhaps warrants revisiting their other findings and recommendations.

This review has identified a number of barriers which have likely held up the generation of a consensus view. These factors included muddled terminology; discrepancies about essential and non-essential features; and theories which either miss or ignore alternative cases thereby limiting their relatability and usability in other contexts. The review has reflected on factors influencing CT development in the modern classroom, explored issues related to a lack of a consensus definition, as well as how the usual culprits of not enough time, training or funding can thwart CT efforts. Hopefully, awareness of these features, will both encourage theorists to be mindful and fastidious when expressing ideas about CT and will assist educators when reading, reviewing and generating their own understanding and approaches for CT.

While ideas concerning the development and assessment of CT are presented in Chapter 4 (Section 4.1), explorations of ideas about the definitions and context for CT as well as educator perspectives on CT, have revealed reasons to expect variety and disparity in this area. It is challenging for educators to make decisions about CT without the benefit of a common definition, data from evidence-based practice, or an understanding of all the various perspective on CT and how they align to the different assessment tools. However, synthesis of the literature for this review has led to the production of a conceptual framework about CT that could help educators with this (described in brief in Section 1.2.4, and continued in Chapter 2). This framework reveals the key processes in CT, and the relationships among them (see Chapter 2 for a more thorough explanation). This framework has been designed to highlight to educators and commentators that it is the product of CT (sound, well-reasoned judgment) which is applicable across perspectives, whilst the process of CT is shaped by the discipline-specific nuances and criteria. However, the purpose of the framework is more than to establish what CT is. It is intended as a tool to show educators how they can incorporate the relevant parts of existing perspectives on CT into their pedagogy, rather than needing to develop or find an approach that perfectly embodies their goals.

It should help them to identify which aspects of CT they are focusing on or missing and this will help to guide their choices when making pedagogical decisions.

This literature review has identified core areas of understanding in relation to the concept of CT. It also briefly explored CT development. Notably, the perspectives of those needing to work at the interface between CT conceptions and CT development (in other words, educators) are largely missing from the literature. Additionally, there are numerous reports that educators are (or feel) ill-equipped to train the next generation of thinkers. In order to be able to produce graduates that are equipped with the thinking and problem-solving skills they need to solve tomorrow's problems, there are both conceptual and educator-related obstacles to overcome. This research intends to address some of the issues by providing conceptual and classroom-based insights into CT. The next two chapters supply conceptual insights that address some of the existing gaps in knowledge identified in this review. Specifically, Chapter 2 supplies a reconfigured conception of CT (the ACT Framework) that unifies the core aspects of CT but creates a frame that is flexible to context; while Chapter 4 explores educator perspectives about the nature, development and assessment of CT. These three chapters form a frame that explores the conception of CT through theory and practice. The remaining chapters then explore CT through the frame of development and assessment from a bottom up perspective to supply new understanding about CT development in Australian science classrooms.

Chapter 2

A New Critical Thinking Definition and Framework

Fitting together the pieces of knowledge into a coherent framework is the art of science.

Lederman, 2008, p.399

2.1 Introduction and chapter outline

The exact nature of critical thinking (CT) has remained somewhat elusive for theorists and practitioners (Lai, 2011), even though there is agreement about CT development being a core goal of education (Facione, 1990a; Pithers and Soden, 2000; Shehab and Nussbaum, 2015; ŽivkoviĿ, 2016). Previous perspectives about CT allude to it being a process, a product and a way of being. CT is captured through various theoretical lenses and as both dependent and independent of learning domains (Sternberg, 1989; Ennis, 1997). Essentially, CT is a complex multi-faceted construct. These ideas were revealed throughout the literature review, which also reflected on the additional complications resulting from the variation, duplication, and bewildering language used to describe CT across the theoretical perspectives. Peirce (1872) expressed that doubt causes an internal struggle that will only cease when one "attains a state of belief" (p6). Yet with so many available perspectives on CT, in addition to a lack of clarity in the language around CT, it is not surprising that doubt remains. This chapter addresses these issues by presenting an adaptive framework that can handle the variability across the CT landscape.

During the course of generating my own understanding of CT, I identified categories to gather and place the various perspectives on CT. The process helped me to understand the range of perspectives on CT, and the common and core attributes associated with CT. But the further I got into this doctoral journey, the more I realised the potential of this organisation tool for helping others to understand CT too. With further refinement, and examination of the patterns of thinking engaged in the CT process I was able to formulate a multi-dimensional adaptive framework that incorporated the shared thinking on CT yet remains adjustable to the points of difference. This framework was briefly introduced in the literature review; however, this chapter is dedicated to describing the framework, its supporting definition and the CT process. This chapter demonstrates how this tool is used as a starting place for understanding CT, or as a means to organise and apply existing thinking about the attributes and processes of CT. It also discusses the potential for the framework to be used to aid curriculum development and pedagogical decisions. Essentially, this chapter demonstrates and explains how this theoretical component of work is a tool that aligns with the underlying goal of my research – to equip educators with knowledge and tools to help them develop CT in their classrooms.

The development of the framework and its definition were outcomes of exploring existing perspectives on CT. This involved collating the common attributes and principles; noting the core points of differences in perspectives; and determining whether they were true differences or just variations in terminology. The next part of the process involved reflecting on the interactions between the common components. This enabled reasoned judgments to be made about which components were both necessary and sufficient for CT, and which components enhanced the CT process but were not essential to the basic form of CT. Lastly, the nature of each of the essential components was re-examined to determine their most basic function within the CT process. This helped facilitate the adaptability of the framework because there is variation in the emphasis and importance of particular attributes within the different conceptions of CT. Stripping the components back to their core function in the CT process sets a minimum standard and function for their role within the framework. It also provided concrete terminology for educators to use to explore the literature. However, the most relevant aspect when focussing only on the core function is the modification of the application for the user's unique context. This meant framework users would be able to decide how much to emphasise each component. They could also determine which of the existing theoretical conceptions best align to the framework components so that the CT process features the important skills and criteria for their context. The result of all this deliberation was a new framework, called the Adaptive Critical Thinking (ACT) Framework, and a supporting definition. The ACT Framework contains five adaptable elements that capture the shared thinking on CT (see Figure 2.1), which can be deployed in various configurations. The ACT Framework definition helps maintain the rigor of the element adaption and deployment. It conveys that CT requires all ACT Framework elements for the judgments to be considered a product of CT (and not just 'thinking' or opinion). Descriptions of each of the ACT Framework elements are outlined in Section 2.3, and some suggestions about how educators can incorporate these general, discipline-specific, explicit and implicit aspects into their classrooms are explained in Section 2.4.

The Adaptive Critical Thinking (ACT) Framework definition: *Critical thinking is a purposeful inquiry process that involves the deployment of thinking skills and context-appropriate criteria to make reasoned, reflective judgments and transform information into useable knowledge to resolve points of uncertainty* (Butler, 2018, 2019).

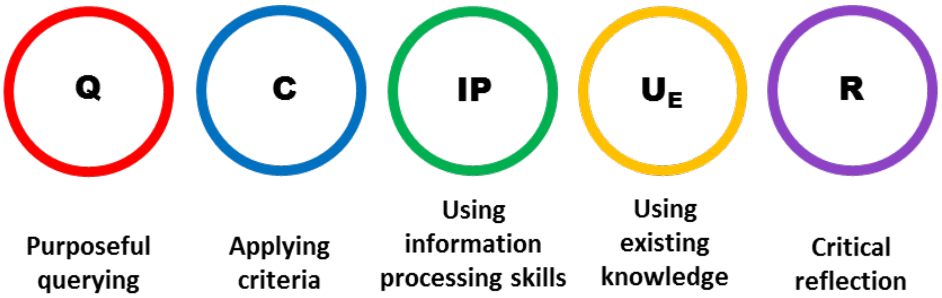


Figure 2.1 Essential elements of the Adaptive Critical Thinking (ACT) Framework. Note the elements are not in any implied order, nor do they interact in a linear fashion. The interaction between elements is explained in Section 2.3.

The ACT Framework emerged out of a sense-making task that involved a reflection on what aspects of CT to incorporate into my survey on Australian educator perspectives. Ironically, it grew from critically thinking about why a pervading belief about the need for a consensus exists when there is so much commentary on CT. The supporting definition emerged from a desire to maintain the ACT Framework's integrity, preventing the partitioning of the elements in the adaptive framework, as has occurred with other models and taxonomy's such as Blooms (Anderson et al., 2001). Facione reported going through a similar calamity (1990b) - the intent behind the Delphi panel was not to generate a definition of CT, but an approach to its development.

The ACT Framework definition reveals CT as a process of transforming information into understanding and knowledge, by investigating the construction of that information and its related components so a judgment can be made. However, the broader function of the definition and framework elements are to provide structure and terminology to the CT process. This process is explained next before discussion returns to the elements and how the definition and framework function as a highly adaptable toolset for guiding thinking.

2.2 Explanation of the CT process

When trying to describe CT, it is necessary to consider what the essential ingredients are, and what can change without ultimately changing the purpose of the process. Dewey (1933) considers CT a sequence of chaining events that move from reflection to inquiry, to critical thought processes that conclude with an evidence-based judgement. More recent conceptions emphasise some of these parts or focus on the standards of thinking. Having so many different perspectives creates quite the challenge for determining the core purpose of CT. However, by incorporating the shared thinking on CT, the ACT Framework and its definition serve to highlight its essential ingredients and characterise the process of CT. The ACT Framework and supporting definition highlight that the CT process is a way of generating targeted understanding about information. It involves exploring the information and making reasoned judgments about the validity of the information (and its construction and connections), with all of the decisions and judgments made shaping how the information is integrated with existing knowledge for future use. Importantly, if a person has employed the CT process whilst thinking about something, they will have an evidence-based understanding of why they know what they know. Potential outcomes of the CT process therefore might be - discard this information because of… {*relevant reasons or justification*}; store this information with these other things I know because of … {*connections and reasons*}; and/or think about the information and its connections some more because I still do not know … {*further ideas or questions for follow up*}. However, regardless of the way the information is judged, the overall outcome increases the understanding of the subject matter and a new baseline of understanding for all further thinking. Essentially this framework and its supporting definition provide a mechanism to understand the journey information takes to become useable knowledge.

This journey is depicted in Figures 2.2-2.4. Figure 2.2 shows the possible pathways that information might travel as a person encounters it. They might use fast thinking (System 1) drawing on their intuition or emotions to guide their decision-making (Kahneman, 2011) and decide to ignore the information or accept it without further thought. Alternatively, they might slow their thinking down (System 2 thinking) to query the validity of the information and run it through the CT process. Importantly, the understanding (and therefore, usability) of the information generated through CT is greater than that from System 1 thinking. Figure 2.3 shows one cycle of CT, as well as how engagement with the CT process can be self-propitiating, leading to the generation of other questions to follow up as information pertaining to the original query is gathered. The conclusion of the CT process means the thinker has satisfied their curiosity (or resolved their current doubts) relating to the initial query at that time. However, sometimes the conclusion of the CT process may relate more to logistical constraints (such as a deadline to plan or complete an assignment) than resolving all doubts about an issue. Finally, Figure 2.4 shows multiple cycles of the CT process, illustrating how ongoing engagement with the CT process constantly changes the thinker's existing knowledge base. Collectively, these figures illustrate how information is filtered and integrated into usable knowledge. They show that the process of CT draws on previously defined knowledge constructs (existing background knowledge) to generate increased understanding of new information. They also show how this increased understanding becomes the new version of existing knowledge for further cycles of CT. This means existing knowledge is both a part of the CT process and changed by the CT process.

The framework incorporates background knowledge through the element 'using existing knowledge.' Background knowledge plays an important role in CT because CT "always takes place in the context of (and against the backdrop of) already existing concepts, beliefs, values, and ways of acting" (Bailin, Case, Coombs and Daniels, 1999, p.290). It is also important because the things we already know can act as a filter for decision-making. For example, after ten years of use, your vacuum cleaner needs to be replaced. You see ads for sales on vacuum cleaners at two different stores that you have been to before, your existing knowledge of the helpfulness of staff and pricing of those stores will play a role in guiding where you go to buy the replacement (or whether to look elsewhere or wait for a better sale). So existing knowledge is an important element in the CT process because it shapes initial perceptions of incoming information and whether it merits further investigation - it can alert you to the value of that information in that particular moment.

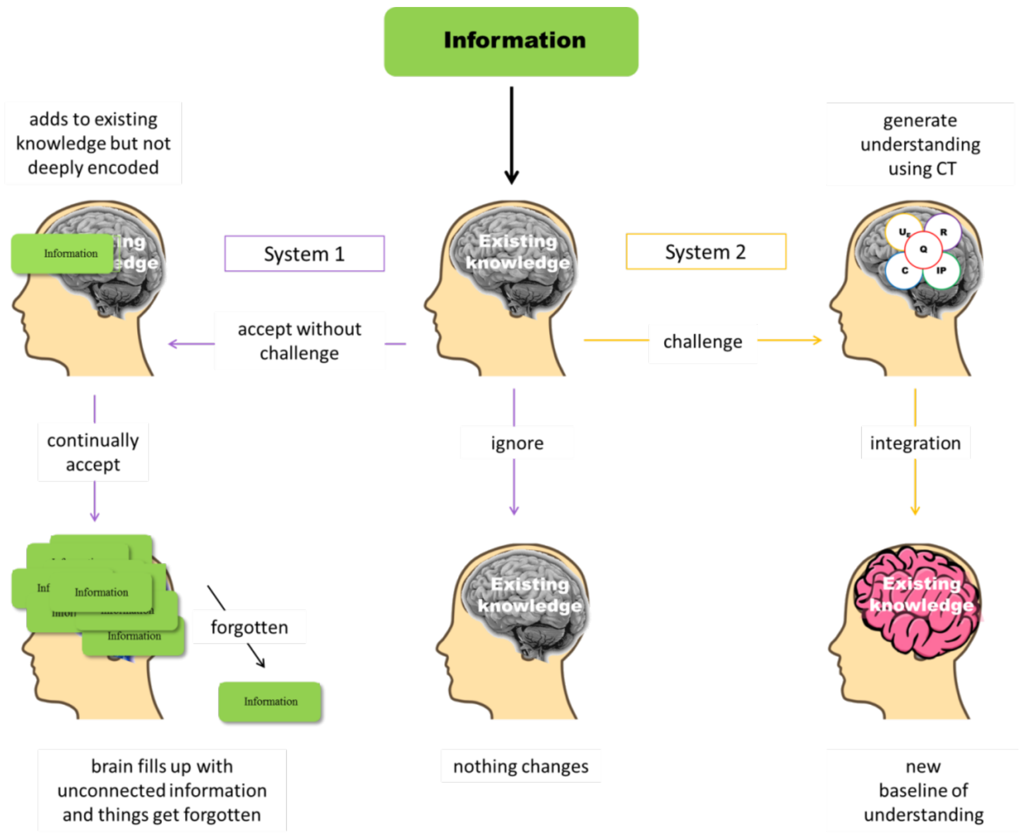


Figure 2.2 Information can flow through different thinking pathways. The circles pictured in the mind at the 'generating understanding using CT' step are the ACT Framework elements. CT challenges the information and generates a new baseline of existing knowledge. Whereas in the other pathways, existing knowledge either does not change, or it grows but not in an integrated fashion, making it easier for things to be forgotten over time. Figure created for Butler (2018, 2019).

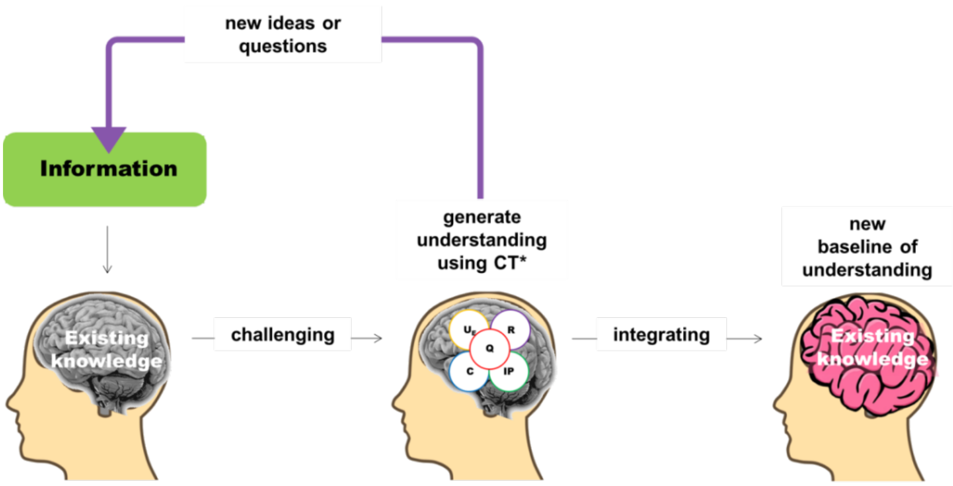


Figure 2.3 One cycle of the CT process shows how challenging knowledge through the ACT Framework elements helps to generate understanding about that information and integrate it so the understanding generated becomes the new usable baseline for further thinking. The purple arrow highlights how undertaking the process of CT can generate new ideas that the thinker may want to investigate. *The amount of time spent generating understanding about any issue can vary with logistical constraints. Figure created for Butler (2018, 2019).

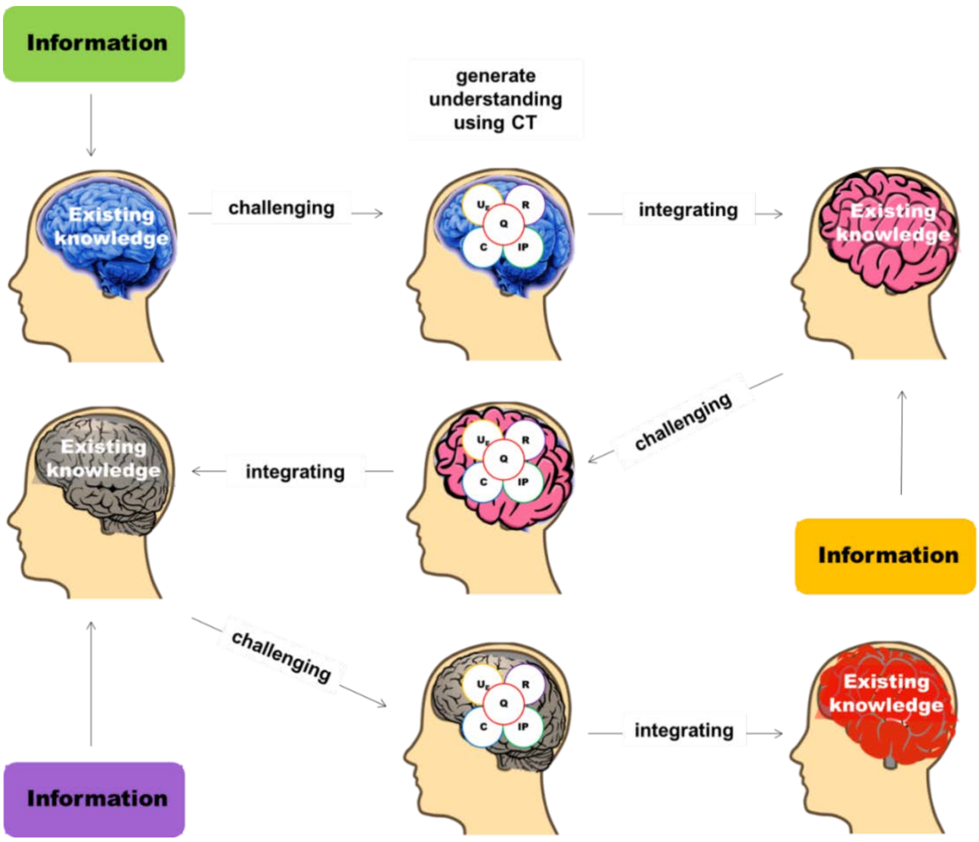


Figure 2.4 Three cycles of the CT process, showing how a critical thinker will consistently interact with new information. At each encounter, the thinker challenges the new information through the ACT Framework elements to generate understanding about that information and integrate it into their existing knowledge. It shows that the understanding generated after each iteration of CT becomes the new usable baseline of existing knowledge for further thinking. Note: instead of undertaking the CT process every time new information is encountered, there may be instances when the thinker opts for one of the other pathways shown in Figure 2.2. Figure created for Butler (2018, 2019).

2.3 Explanation and application of the ACT Framework elements

Analysis and synthesis of the various perspectives in the literature led to the discovery that CT is an inquiry process that commonly includes five core cognitive actions. These include purposeful querying, critical reflection, applying criteria, using existing knowledge, and using information processing skills. These framework elements are explained in Table 2.1. The elements serve as anchor points for users to connect their existing understanding of (and future understanding about) CT, whereas the definition directs their collective application. For example, in this new model, a reasoned judgment requires the activation of all the framework elements. In isolation, each framework element is not sufficient to be CT, just as "carrying out a set of procedures is not sufficient… since any procedure can be carried out carelessly, superficially, or unreflectively" (Bailin, 2002, p.363). If the thinker performs just some of the parts, perhaps taking out the validation components (applying criteria and using critical reflection), then what remains is uncritical thought in the form of a question, an idea, or opinion. For example, judgment without reasoned criteria and information processing is just an opinion; using information processing skills without a query or criteria is just thinking etc. This is why the elements included in the ACT Framework are described in verb form because CT emerges from using these things in concert. However, it is worth noting there are multiple ways that the elements can interact and prompt further questioning because the elements interact like strands of a rope rather than in a linear fashion. This means the thinking done in one aspect might change the course of the thinking done in another.

For CT to be able to derive usable knowledge based on deep understanding that the thinker can draw from for future thinking, every cycle of CT should include:

- Generating or responding to a question, problem or something the thinker has noticed about a piece of information.
- Using information processing skills for gathering and processing extra information about the construction of the original information, and potentially other sources to verify the original information.
- Judging the original information (and any extra information gathered about it) against criteria. The criteria should assist with forming understanding about the issue with an appreciation of how arguments are justified at an appropriate standard for the context.

- Reflecting on the information and judgments to seek for issues in the construction of the original information, and/or to seek for issues in the construction of the new understanding about that information and/or to check that the purpose of the CT has been satisfactorily addressed.

However, just as the explanation of the elements is to supply the basic form of CT, this description of the application of the elements is a very general description of what the process involves (see Section 2.3.1 for a more complete description of how to apply the elements and Section 2.3.3 for explanations of how to adapt the elements).

Table 2.1 Explanations of the elements in the Adaptive Critical Thinking (ACT) Framework.

Element	*Explanation*
Q **Purposeful querying**	This element involves **identifying or engaging with a point of difference or a problem to investigate**. It is included in the framework to differentiate thinking from CT, because CT is always goal or purpose-driven, whereas thinking does not have this requirement.
C **Applying criteria**	This element involves **applying rules or constraints to information**. Criteria help guide decision-making about the importance and/or validity of the information, in order **to provide contextually appropriate judgments**.
IP **Using information processing skills**	This element involves **using CT skills*** such as those outlined in Facione (1990a) or other theorists **to explore the problem and help resolve the points of doubt**. However, this element is titled information processing skills rather than CT skills to remind educators/users that whilst thinking skills are necessary to do CT, not all thinking skills are sufficient to be described as 'CT', because thinking in isolation from purpose and weighted judgments is just thinking or information processing, not CT).
U_E **Using existing knowledge**	This element involves **filtering/ examining information through previous knowledge/mindsets**, as the thinking that we do is shaped by the thinking that we have already done and influences how we approach current and future problems.
R **Critical reflection**	Critical reflection involves **quality checking thinking,** but it is important to note that this thinking is not limited to a self-reflection and introspection. This could involve scrutinising **your own thinking**, **and/or** it could include quality checking **other peoples' thinking**, for the purpose of identifying biases, flaws in logic and noticing gaps in thinking.

*CT skills commonly agreed upon include analysis, inference and evaluation (Lai, 2011).

2.3.1 Additional information about applying the elements

There are a few fixed rules about the application of the elements. However, CT is a process of applying criteria and judgment to information with a specific purpose - be it to resolve a problem or doubt about an issue, or assist with making a well-reasoned decision. This purpose is the main feature that distinguishes uncritical thinking from critical thinking. Because of this, **the application of the ACT Framework always starts with some purposeful querying** to set the tone for the inquiry. However, the next element or elements applied during the CT process will depend on the nature of the question. It could involve reflecting on the information and existing knowledge, applying criteria to the information, or gathering more information or a few of these in concert. The only other requirement of the ACT Framework is that there is **at least one iteration of all the elements**, as this helps to ensure the thinking generated is sufficient to be CT. See below for some further explanations to assist with applying the elements. However, more complex details about adapting the ACT Framework, such as incorporating domain-specific or domain-general aspects of CT into the framework as well as ideas about explicit or implicit teaching for CT, are addressed later in this chapter.

Purposeful querying

Critical questioning is a vital part of the construction of knowledge in the CT process (Badia, 2016; Mayweg-Paus, Thiebach and Jucks, 2016). Application of the purposeful querying element might involve starting with a pre-determined question about the initial information, or it might be a two-part process. In the two-part process, the thinker might ask themselves if they trust this new information, or the thinker might notice or observe something about the initial information that does not fit with their existing understanding of that issue. The latter can result in the generation of a more specific question that will set the approach for the inquiry. However, more questions may be generated as the CT process proceeds.

An additional point to consider is that when generating a question about the initial information, then a few of the elements will work in concert in the initial stages of the CT process. For example, generating a question can involve consciously comparing the initial information against the thinker's **existing understanding**, or intuitively noticing that something about that initial information does not align within existing understanding. It could also involve briefly exploring the construction of the initial information and then **applying criteria** to that extra information to determine whether the thinking and/or methodology that led to the generation of the initial information is appropriate and rigorous for that context.

Using existing understanding

Through the CT process, the initial information is questioned and compared against existing understanding. When applying this element, if something different to the thinker's existing knowledge is noticed about the initial information (or the construction of that information), then a query is generated. The generation of such queries can occur at any time in the CT process. Existing knowledge, be it about content (domain knowledge), or general problem-solving processes, is important for information retrieval and searching (Hirsh, 2004). It is also important for providing a basis to reflect on unknowns and unpack a question or query (Pintrich et al., 1993). Since any new thinking is always done in the presence of existing understanding (whether consciously or subconsciously), this element is effectively always 'on' in the CT process, and consequently **often occurs simultaneously with** the application of **other elements**. Awareness of the role and function of existing knowledge in the CT process reminds educators to be mindful about the complexity of student's understanding. Students need to be able to form connections about why they know what they know as they learn so that the depth of their understanding forms a strong basis for doing CT.

Applying criteria

Criteria are "reliable" reasons (Lipman, 1987, p.40) that help to "govern critical deliberation and judgement" (Bailin et al., 1999, p.291). The role of the applying criteria element in the framework is to provide rules and standards to the thinking, by supplying tools for judging the validity of information and making weighted comparisons of the information. However, there are a few things to know about criteria before considering how to apply this element.

Firstly, there are criteria for helping form a question; there are criteria for helping answer a question to a specific standard; there are criteria for checking that an answer is at an appropriate standard for the context. These may not be the same. Secondly, criteria can be generic or more broadly applicable, such as Paul and Elder's (2006a) intellectual standards and traits for good thinking. Conversely, criteria can be specific to a context, supplying the rules that govern the mode for high-quality thinking in that specific situation or context. These context-specific criteria could relate to the scope of the inquiry, the research design or experimental methodology, or could include parameters around managing or presenting subject matter and undertaking data analysis and interpretation. For example, when making decisions and judgements in science, more weight is placed behind quantitative than qualitative evidence; whereas in disciplines such as history, there is a different appreciation of qualitative evidence. However, criteria do not always have to consist of domain-general or domain-specific rules; they could also be logistical constraints (such as limited time, limited money, and/or near a specific location or facilities).

Knowledge of criteria and the standards for high-quality thinking can be useful for helping thinkers deal with unfamiliar problems. For example, a thinker may be presented with a situation where their specific existing knowledge may not assist them, but considering the likeness of the problem to other problems they have solved will help them make better decisions about how to proceed with the new problem.

The main aspect to understand when applying this element is that knowledge of appropriate criteria for high-quality thinking and decision-making is essential for the rigorous application of this element in CT. However, not everyone will immediately have the **existing knowledge** or expertise needed to achieve this standard of thinking in the context they are making the decisions or judgments. In an education setting, the educator could scaffold ideas relating to the applying criteria element to move students toward a stronger understanding of what appropriate criteria are, and also how to apply them. This element strengthens as the learner's ability and understanding develops, and the skill with which this element is applied increases with experience. Outside of an education context, if the thinker does not already have an awareness of the relevant criteria, then they need to seek it out (using the **information processing skills** element).

Critical reflection

The extent to which a thinker engages with the critical reflection element will vary. As long as there is some reflection about whether the thinking and understanding generated by the application of the other elements has satisfactorily resolved the initial query, then this element's basic role is fulfilled. However, the depth of understanding is enhanced if this element is also applied by spending time quality checking the original information. This quality checking will involve **gathering information** about the construction of the knowledge and **applying criteria** to these **evaluations** of others thinking. This element can also be applied to quality check the thought processes generated throughout the CT process (one's own thinking), however, this type of metacognitive checking relates more to enhancing the CT process than meeting the minimum requirements for thinking to be CT. The value of thinking about the quality of one's own thinking is explained in Section 2.3.3.

Using information processing skills

The application of the information processing skills element is to help the thinker generate further understanding of the original information. Using this element to seek out additional information about the original issue, helps the thinker to verify or discredit the original information. There are many different thinking skills associated with CT and information processing, so the thinker needs to choose which thinking skills to deploy to gather and process information that will best answer the initial query. The application of this element may occur concurrently with **applying criteria** and **critical reflection**, as these aspects can bring structure to decisions and awareness about the thinking skills needed to address the problem. When this element is applied alongside the applying criteria element, the purpose of applying information processing skills is to gather specific information (or to refine a set of information). When the information processing skills element is applied without the applying criteria element, the purpose of the skills would still be to expand understanding by gathering information, but this process would not have any specific constraints; or it could be to refine a set of information without any pre-defined constraints. If a framework user is looking for some additional terminology to seek for this element, information processing skills are representative of the skills listed in Facione (1990b) and also to higher-order thinking skills and information literacy skills (see Reece, 2005).

2.3.2 Examples of the activation of the ACT Framework elements

Since CT is a universal skill with context-specific nuances, there are many different ways that the CT process might be prompted or proceed. Figures 3.5-3.8 show a few general examples of the activation of elements throughout the CT process. The order the elements activate varies depending on the nature of the inquiry, and the level of experience the thinker has with the CT process. Figure 2.5 depicts the element flow when the thinker forms the question. Figure 2.6 shows a possible flow of elements when the question has been supplied to the thinker. These first two examples represent the minimum set of thinking processes that would be undertaken when doing CT. Another feature to notice about these first two examples is that the critical reflection element is only applied once, and at the very end. However, there can be value in applying this element sooner and more often, to help with the identification of flaws in the thinking process. Figure 2.7 shows the extended CT process where the first round of applying the elements did not resolve the initial query because somewhere in the thinking process, there was a flaw in the thinking. In this example, the critical reflection element was again only applied to help confirm that the newly generated understanding addressed the initial query to an appropriate standard.

However, the initial query was not resolved. Therefore, the thinker now has two queries to address, and additional critical reflection is needed. The thinker needs to determine what went wrong in their initial CT process, and also what additional understanding is needed to address the initial query. This requires the application of the critical reflection element in a metacognitive fashion, with the specific purpose of reviewing the thinking done at each stage, to identify flaws in the thinking process and help resolve them. Figure 2.8 shows an example of a thinker who applies the critical reflection element at multiple stages in the CT process. The use of the ACT elements in this manner is representative of increased experience with the CT process. For example, the purpose of applying the critical reflection element as per Figure 2.8 is to help monitor the thinker's thinking and identify flaws in the CT process sooner than leaving it to the stage of checking if the new understanding sufficiently addresses the question. This helps to circumvent arriving at what could be the end of the CT process only to discover that at some point during the process, something went wrong. Whereas waiting until the end of a cycle can result in the need to continue to do CT to answer the initial quest and to recheck the thinking steps to identify where it went wrong so the thinker can avoid this mistake in future CT. However, with all of these examples, and really at any point in the CT process, the thinker can choose to stop undertaking additional thinking and reflection. Stopping early still results in increased understanding; however, the full query is unlikely to have been resolved.

Another important aspect about the CT process emerges when reviewing Figures 2.5-2.8 collectively. These figures demonstrate there is a lot of flexibility in the order of the stages of the CT process, and also show that both the starting point and ending point can be prompted or self-initiated. The flexibility of the elements increases the potential applications of the framework. It is possible to direct both the contextual framing of the framework elements and the order that the elements engage. The potential for variability in engaging the elements through the CT process is demonstrated through Figures 2.5-2.8. However, what was not previously explained is that the role each of the stages plays varies depending on the situation and context of the problem. This is why some of the steps are duplicated or revisited during the CT process. Figures 2.6 and 2.7 displayed examples of situations where the thinker is supplied with a question. These figures are representative of the CT process in an education-setting context, where an assignment question may supply the problem to direct the student's thinking. In an education-setting context, the 'purposeful querying step' can result from an external demand to identify the criteria to address the question presented, and this can result in a duplication of the 'applying criteria' element.

Students may consider the criteria prior to engaging in information processing, to structure the thinking toward relevant ideas, and the 'applying criteria' element may also be employed to confirm that they have addressed the outlined problem (see Figure 2.6). However, there is a subtle difference in the criteria step between the first and second iteration. In the first instance, identifying the criteria becomes a part of the purposeful querying step. However, in the second instance, the criteria check arises through reflective thinking (critical reflection element), as the student considers the thinking previously done in the topic and whether the query has been resolved or if there are gaps in their thinking. Identification of a flaw in the thinking, or recognising that they have failed to answer the assignment question satisfactorily, may prompt students to reflect on their previous thinking to identify where it went wrong, and to generate additional understanding (see Figure 2.7). If a student fails to identify flaws in their thinking, the feedback received in an education-setting may also help a student realise they have failed to address their initial query satisfactorily. As students become more experienced with the CT process, they will become more proficient at using the applying criteria element and in recognising flaws in their thinking.

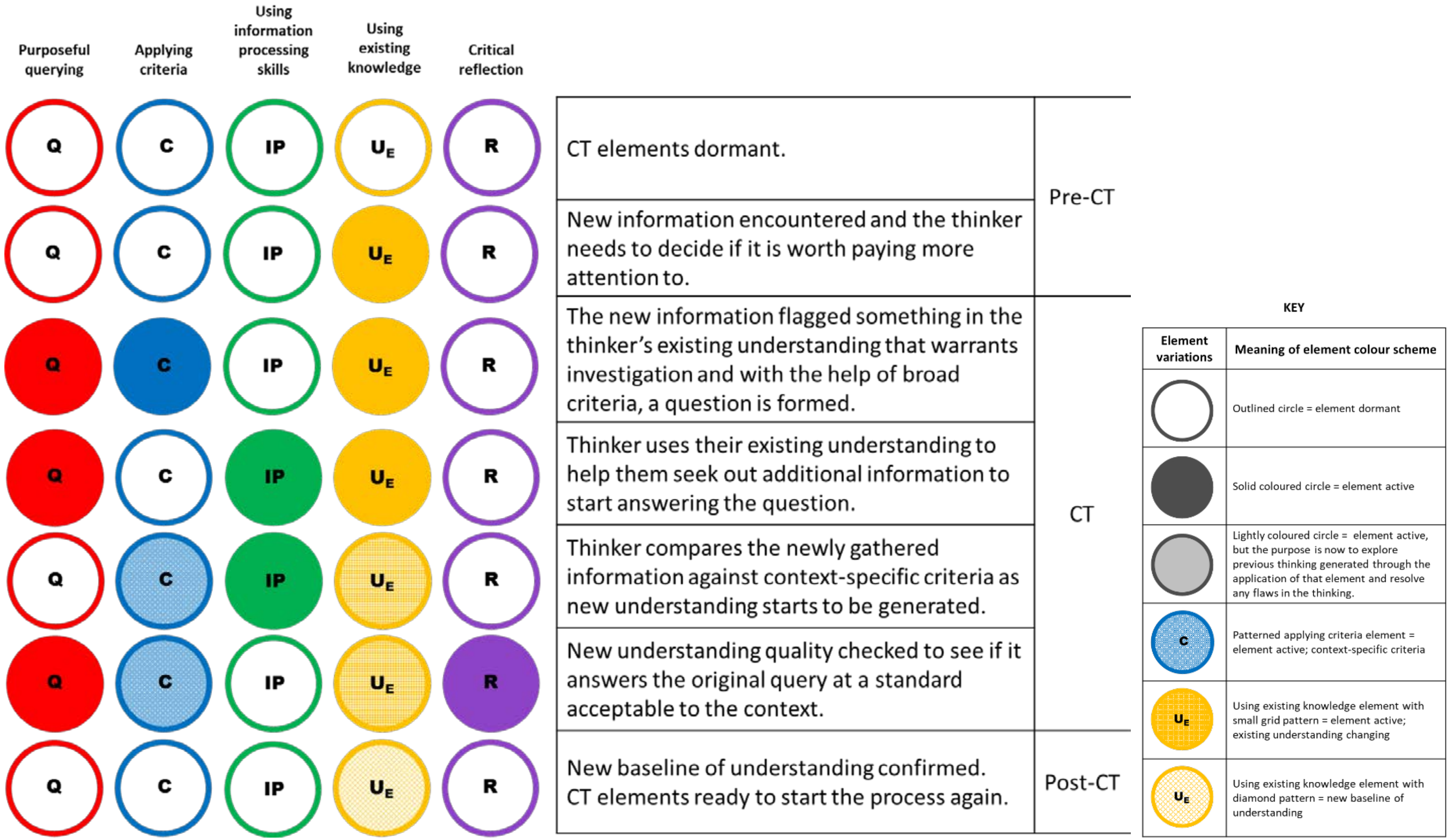


Figure 2.5 Engagement of elements when the thinker forms the question. In this example, the critical reflection element is only engaged at the final stage of the process (this is the minimum expectation). However, this element could also have been engaged earlier in the process, such as with the information processing element, to put some focus on quality checking the thinking that went into the construction of the original information. Figure created for Butler (2019).

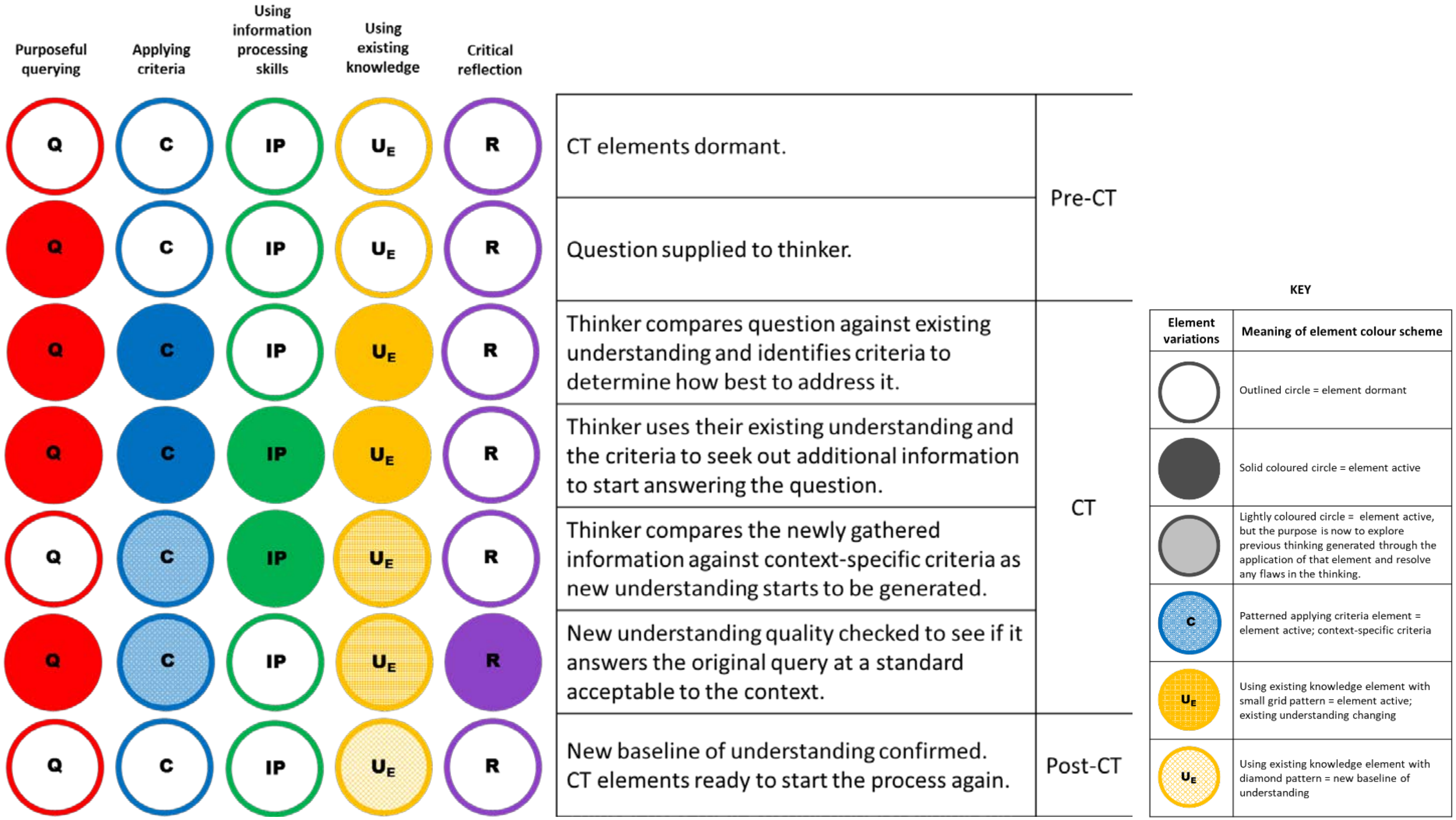
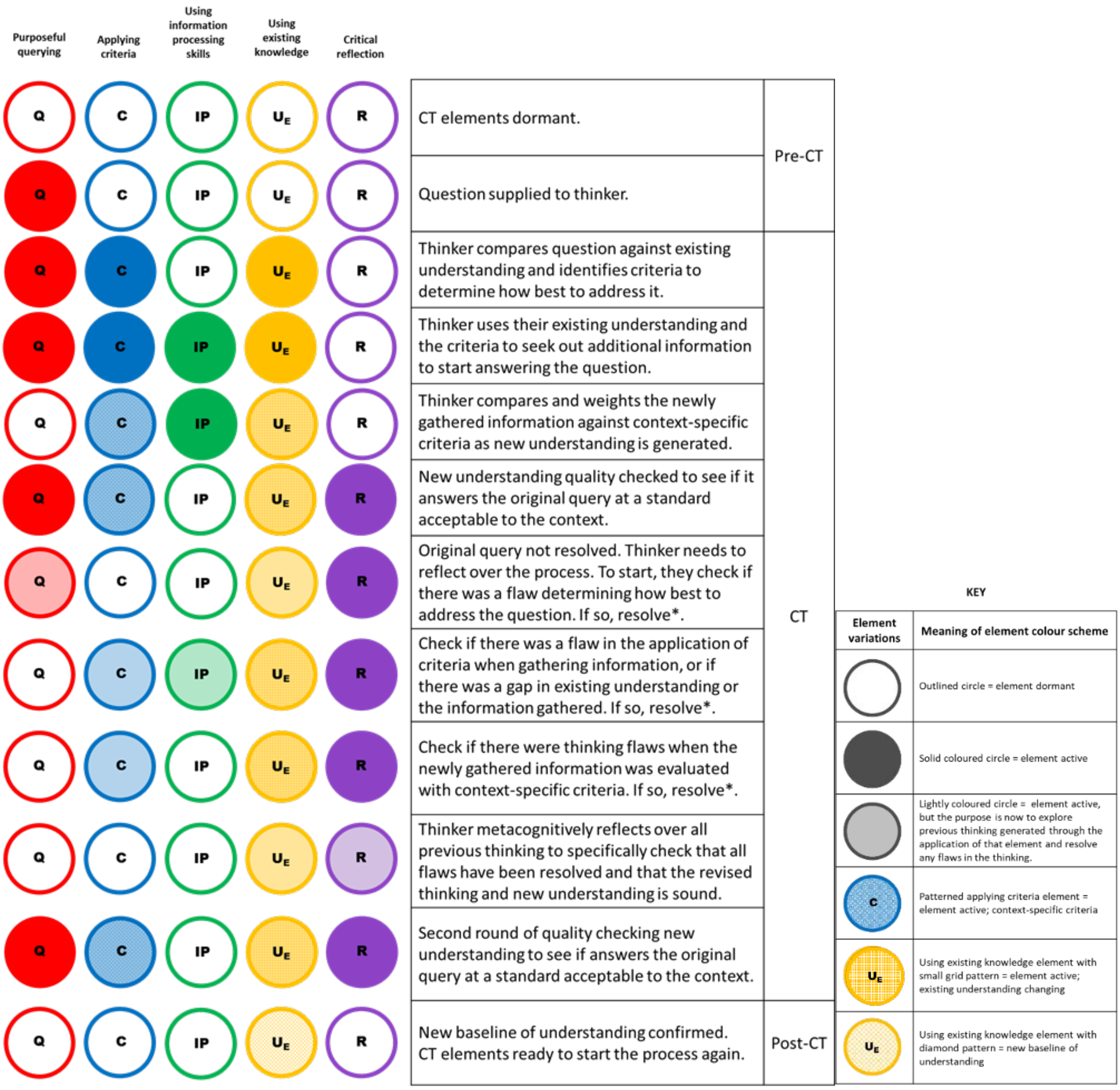
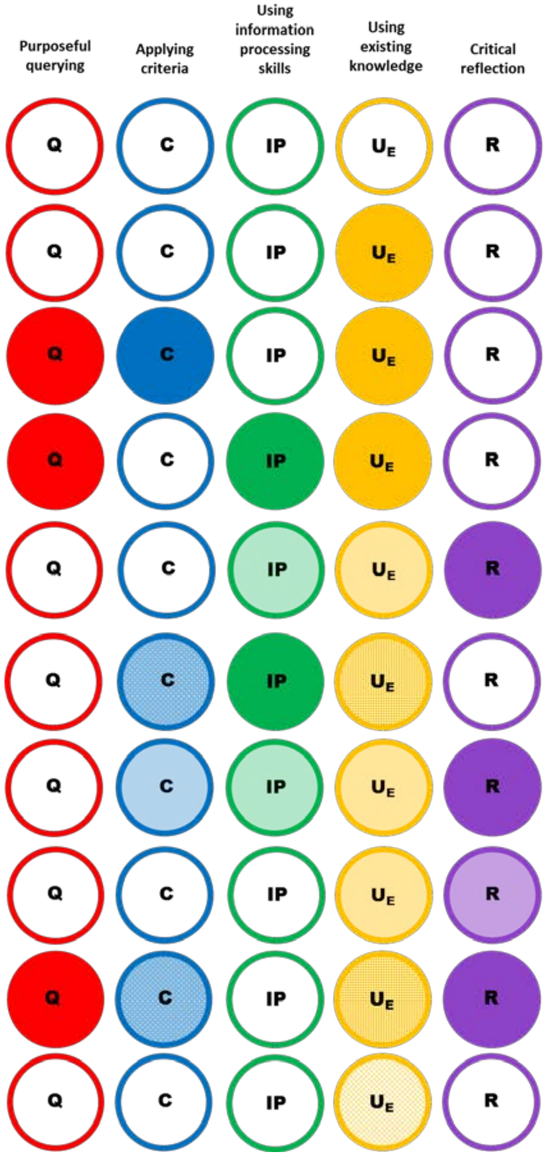


Figure 2.6 Example of engagement of the elements when question supplied to the thinker. In this example, the critical reflection element is only engaged at the final stage of the process (this is the minimum expectation). However, this element could also have been engaged earlier in the process, for example, when examining the question and determining the criteria needed to address it. If the critical reflection element was engaged there, it would remind the thinker to be mindful of the standard of the information they gather using information-processing skills. They may consider things like - is the information from a trusted source? Could it be biased? Figure created for Butler (2019).



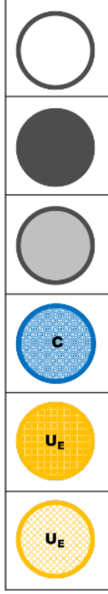
*Once flaw has been identified and resolved, thinker can resume standard element application, or they can continue to apply elements with a reflective thinking process.

Figure 2.7 Example of engagement of the elements when a query is not resolved after one iteration of all CT elements. In this example, the thinker was supplied with a question; however, the same flaws could occur if a thinker generated their own question. In both cases, the flaws could have been circumvented if the critical reflection element was engaged earlier in the process and/or throughout the process. Figure created for Butler (2019).



CT elements dormant.	Pre-CT
New information encountered and the thinker needs to decide if it is worth paying more attention to.	
The new information flagged something in the thinker's existing understanding that warrants investigation and with the help of broad criteria, a question is formed.	
The thinker uses their existing understanding to help them seek out additional information to start answering the question.	
The thinker reflects over the thinking done at the last step to check if there are any additional ideas to investigate, and/or if there are methodological issues with the information they have collected. They may also check for flaws in their own thinking.	
Thinker compares the newly gathered information against context-specific criteria as new understanding starts to be generated.	CT
Thinker reflects over the thinking done at the last step to check if there are any additional criteria to consider, and/or if they have been applied at an appropriate standard for the context.	
Thinker metacognitively reflects over all previous thinking and the new understanding to do a final check for potential flaws in thinking processes which generated the new understanding.	
New understanding checked to see if it answers the original query at a standard acceptable to the context.	
New baseline of understanding confirmed. CT elements ready to start the process again.	Post-CT

KEY



Element variations	Meaning of element colour scheme
	Outlined circle = element dormant
	Solid coloured circle = element active
	Lightly coloured circle = element active, but the purpose is now to explore previous thinking generated through the application of that element and resolve any flaws in the thinking.
	Patterned applying criteria element = element active; context-specific criteria
	Using existing knowledge element with small grid pattern = element active; existing understanding changing
	Using existing knowledge element with diamond pattern = new baseline of understanding

Figure 2.8 Example of the CT process with the activation of the critical reflection element multiple times throughout the CT process. In this example, the thinker generated their own question. Figure created for Butler (2019).

2.3.3 Adapting the ACT Framework elements

CT is dynamic and improves with experience. Scaffolding and adapting the elements should be structured to allow the thinker to grow in their understanding of the CT process as they gain an understanding of other subject matter. The application of the ACT Framework elements to information demonstrates how CT is transactive thinking that increasingly changes the thinker's understanding, as new information is gathered, processed and placed. Together the definition and framework elements work to ensure that the CT process shifts information from being accepted at face value, to being challenged or tested against criteria to determine the merit or truth of the information. As the thinker increases in experience with undertaking CT, the way they engage with the elements will change, as will the way they think about and view the world. This accumulation of useable knowledge gained through practice at CT highlights why the CT process should be fostered and developed as a part of ongoing learning, rather than delivered in isolation or as a single unit. However, there is further layer to CT. When creating the ACT Framework, it was noted that specific details about the defining attributes related to recognising purpose and sound judgment, as well as the expertise needed to know and appropriately apply the criteria, would vary depending on the context. Even if the core CT processes are the same in domain-general and domain-specific thinking, the details that make the newly generated understanding and judgments sound are context-dependent. In other words, the nature and context of the question will also shape the nuances with which the elements are applied.

To achieve 'reasoned thinking' at a suitable standard, contextual expertise or boundaries are needed to shape the appropriate method for proceeding through these thinking and testing stages. Because of this, both the definition and framework leave the details about information processing skills and criteria open to user's preferences about domain-general or discipline-specific nuances, and explicit or implicit expressions. This enables the application of the framework elements such that the thinking is appropriate to the context (the learning domain, general life decisions, etc.). Further, the application of the adapted CT elements means the transformation of the information will produce contextually relevant useable knowledge that is also at an appropriate standard. This adaptability is what makes the framework ideal for use in education since it is capable of accommodating the various literature perspectives, and contextual nuances. Table 2.2 highlights how educators can adjust the component elements to incorporate contextually relevant CT skills from the literature and policy documents into their development and assessment plans.

Table 2.2 Suggestions for contextual adaptions to elements in the ACT Framework.

Element	*Adapting the element*
Purposeful querying	This element can incorporate both explicit and embedded teaching approaches, through varying language and how the problem is constructed. Educators can provide questions; guide students through the process of generating questions* or determining the constraints of a problem; or create a culture of student-driven noticing and challenging knowledge.
Using information processing skills	Information processing skills encompass a wide range of thinking skills which can be tailored to the context (can be broad or can draw out discipline-specific nuances). This means this element is readily adaptable to incorporate views about important CT thinking skills held by the thinker, the education system or the discipline they are applied in.
Applying criteria	This step is where domain-general and domain-specific rules or nuances can be added to the framework. Criteria can be used to shape the thinking, so the judgments are weighted by, or based on, standards appropriate to the context. This step is also where educators can explicitly discuss and define CT criteria, or students can be allowed to derive criteria for themselves. Criteria are shaped by internally constructed criteria such as personal morals, beliefs, values and existing understanding; or external pre-defined constraints or rules that may also state how to weight specific properties over others, rather than deciding these things for oneself.
Using existing knowledge	Since the level of existing knowledge is unique to each thinker, the foundation of the element is not adaptable to accommodate aspects of CT theory. However, the baseline of one's existing knowledge changes with learning and CT. Therefore, teachers may wish to check students' existing understanding for gaps in knowledge that they would expect them to have. If a gap is identified, they could scaffold particular content or skills for students to learn so that they can draw from these in future CT.
Critical reflection	Educators may wish to adapt this element to further emphasise self-reflective attributes and the processes to quality check one's own thinking. They might consider aspects such as considering whether one's thinking is logical and if there are other possible outcomes or conclusions. Then applying these same reflections to other's thinking (what logical fallacies, cognitive biases might be present in the perspectives they put forward).

*Mayweg-Paus et al. (2016) have an effective description of critical questioning which educators might find useful because they provide an outline of the areas where question formation and argument structure can fail.

2.3.4 Prompting and enhancing the elements

The main elements in this framework are included because they are common in most perspectives on CT and because without them, a soundly judged CT product is not generated. In other words, these core elements are essential to the process of **doing** CT. However, there are some refining mechanisms that make the CT process more rigorous and directed. These are metacognition and dispositions; both are essential to the traits of **being a critical thinker**. I distinguish here between the notion of **doing** and **being** because metacognition and dispositions have received much attention in the literature, and failure to acknowledge their connection to CT has the potential to lead to this framework becoming another polarising (rather than unifying) perspective. Theorist's perspectives on the value of these components for high-quality CT are relevant. However, I do not see them as essential for doing CT, because it is possible to generate well-reasoned, useable knowledge without self-initiating the process or reflecting on the quality of one's own thinking (see Figure 2.5 and 2.6). The **core elements** in the framework for doing CT are essential because they have the shared purpose of **addressing uncertainty and generating understanding.** Whereas the **enhancing components,** such as dispositions and metacognitive skills, help with the application of the core elements and **keep the individual's thinking processes in check.** To extend the argument, consider the following: -

Dispositions have not been included as a core element because it is possible to do CT without having the natural, habitual inclination or qualities of the mind to do so. For example, education settings may coerce a student into undertaking CT as part of their studies. Schmaltz et al. (2017) note that "the general consensus is that while dispositional traits may play a role in the ability to think critically, the general skills to be a critical thinker can be taught" (p.1). This highlights that something about the nature of dispositions is different from thinking skills. In addition, criteria may be used to remind thinkers to employ dispositions such as staying open-minded, fair-minded, seeking reasons and respecting intellectual authority, therefore, the thinker does not have to have the natural inclination to do so when doing CT.

Metacognition is harder to justify as a non-essential component. In fact, metacognitive thinking was included in Figures 2.7 and 2.8 (refer back to Section 2.3.2) because there is value in the thinker reflecting on their thinking during CT process to identify (or prevent the generation of) flawed understanding. Consider the following. Metacognition is generally defined as thinking about one's own thinking (Tarricone, 2011), as opposed to thinking about subject matter. Yet the reflection done in the CT process does not have to include thinking about the flow and quality of your own thinking processes for the useable knowledge product to be generated (refer back to Section 2.3.2 Figure 2.5 -2.6). The main reason critical reflection rather than metacognition has been included as the core element in the framework is because metacognition can be undertaken without applying criteria or forming judgments or conclusions. But how is critical reflection different from metacognition? The main distinction is the subject of the thinking. Critical reflection refers to thinking about the thinking on the subject matter rigorously, whereas metacognition is thinking about the thinking to quality check the thinking. The thinking about thinking process for metacognition does not require a specific resolution for it to be metacognition. It is possible that metacognition can be done uncritically (Lipman, 1987). This difference in subject matter and outcome of the thinking is the other main reason critical reflection rather than metacognition has been included in the ACT Framework. Critically reflecting on information and understanding to be able to identify gaps and form connections is essential to making appropriate judgments about the subject matter to resolve the initial query (refer back to Section 2.3.2 Figure 2.5). Whereas metacognition brings understanding to the thinking process, but not understanding of the subject matter, because the purpose of the metacognition is to evaluate and enhance the thinking processes and not necessarily generate the final solution to the initial query.

While not core element framework elements, there are a number of ways metacognition and dispositions enhance the CT process and support the deployment of the core CT elements. Metacognition regulates and refines the thinking stages (refer back to Section 2.3.2 Figure 2.7 and 2.8). When undertaken with the intended outcome of improving understanding about thinking done in CT, it becomes almost its own CT loop within the CT process. These thinking checks, in addition to dispositions such as open-mindedness and fairness, may help prevent the thinker from succumbing to certain cognitive biases as they undertake the process of CT. It can lead to the identification of flaws in the thinking process, which means metacognition can help the thinker self-identify issues in their decision making and judgments, rather than discovering this after the fact through a negative outcome or negative feedback. Therefore, those thinkers who incorporate these components into their CT processes will be more capable of producing a higher standard of thinking, in a shorter period of time, than someone who does not employ these thought regulating skills. Metacognition assists the progression through the stages by prompting thought-checking and reviewing logic and biases in one's own thinking. A thinker can do CT without checking their thinking (See Figure 2.5), but there may be flaws in their thinking that they may not notice (but others will). Alternatively, they may eventually notice, but not until they attempt to arrive at a final judgment and realise there is still uncertainty (see Figure 2.7), meaning they need to do some more CT about the matter. Whereas, if metacognition is used to check the thinking at each stage of the CT process (see Figure 2.8), then sounder outcomes can be achieved more quickly than if the thinker waits until after completing the thinking cycle (and then realising they still have not arrived at a sound judgment or addressed the initial query to satisfaction).

Dispositions such as inquisitiveness, self-confidence, diligence and persistence in seeking relevant information, open-mindedness, and fair-mindedness as outlined by Facione (1990a) are an enhancing component because they can facilitate persistence and progress through the CT stages. Dispositions can improve the application of the purposeful querying element because they help maintain eagerness for reliable information and other good habits of the mind. For example, critical thinkers have an “alertness for opportunities to use CT” and a “willingness to” rethink “views” if “reflection suggests change is warranted” (Facione, 1990a, p.13). Dispositions can enhance the applying criteria element by helping the thinker to keep an open mind when considering how to weight information (critical thinkers display “fair mindedness in appraising reasoning” and “prudence in suspending, making or altering judgments” (Facione, 1990a, p.13). The dispositional attributes are responsible for assisting students to effectively organise and progress through the components when the CT is self-initiated (critical thinkers have “clarity”, “orderliness” and “care in focusing attention on the concern at hand” (Facione, 1990a, p.13). They may also link to the ability and propensity to undertake metacognitive checks throughout the CT process.

Dispositions enhance and keep CT processes in check because they help with self-initiation, task attention and motivation, and can help the thinker to remain inquisitive, curious and thirsty for understanding in the face of uncertainty. Metacognition keeps the CT process in check by alerting the thinker to flaws in their thinking as they reflect on the quality of the application of each element. Table 2.3 provides some further suggestions about how the motivation to apply the core elements can arise from external prompts or through internal initiation (driven by the enhancing elements).

Table 2.3 ACT Framework element explanation, which considers how internal and external factors might enhance or prompt the application of elements.

Element	*Thinking step may be driven by*
Purposeful querying	• curiosity, metacognition or learned disciplinary thinking which leads to the generation of questions and ideas to resolve a point of difference or doubt. (internal initiation) • having to make an optimal choice between two or more options, given fixed parameters or resource constraints (time, money, stock availability, political policies, etc.), or that present a moral or ethical dilemma. Sources of conflict can come from education, the workplace or life. (external prompt)
Using information processing skills	• some metacognitive thinking about which thinking tools are needed to investigate the conflict and also some idea generation about the connections between various pieces of information. (internal initiation) • educational, workplace or life prompts that demand a particular thought process in order to solve the problem, such as being told to gather information on a subject to increase knowledge about a new product or technique; or needing to evaluate and identify gaps in existing understanding to improve understanding or uncover potential problems. (external prompt)
Applying criteria	• wanting to make the best possible decision about an issue. (internal initiation) • being told to solve a problem with certain constraints. (external prompt)
Using existing knowledge	• subconsciously engaging when presented with new information, or be purposefully or engaged through metacognition when reflecting on unknowns, and unpacking a question or query and deciding how to proceed. (internal initiation) • being asked to compare existing knowledge to new information or problems. (external prompt)
Critical reflection	• reflecting on one's thinking through metacognitive approaches such as to check if the thinking is sound, or if all the possibilities have been examined, or by considering how/where a new piece of information fits with one's existing understanding about that thing. (internal initiation) • a response to an educationally derived demand (being asked to reflect on thinking); a workplace demand (procedure or product failed and need to figure out why); or a life demand (financial or relationship conflicts that should be resolved). (external prompt)

2.3.5 Some examples of how existing ideas can be aligned to the ACT Framework

The ACT Framework simplifies the CT process and elements into minimum form. This was to enable existing ideas about CT to be aligned to the elements, rather than generating new ideas or fixed ideas about the elements. This was important because new or fixed ideas would render the framework irrelevant to some theorists. For example, the Delphi panel's ideas about the skills associated with CT (interpretation, synthesis, evaluation, analysis, inference, explanation - Facione, 1990a, 1990b) fit into the information processing element, evaluation also fits into the criteria element, and the skills assigned to self-regulation align with certain aspects of the framework's critical reflection element. It is possible to do the same with the learning objectives from Bloom's taxonomy (Anderson et al., 2001) whereby these objectives become the expected skills used when applying the information processing skills element.

Paul and Elder's (2006a) expressions of CT can also be readily incorporated within the framework elements. Their 'elements of thought' component falls under the information processing skills, and their 'intellectual standards' component would be incorporated within the applying criteria element. Paul's intellectual traits could partially be incorporated through the critical reflection element, with any other important traits being turned into criteria for the applying criteria element.

McPeck's ideas (1981) about discipline-specific criteria fit into the 'purposeful querying' and 'applying criteria' elements, where the disciplinary context can give external directives and strategies applied to the thinking process. Similarly, ideas about the general nature of CT (Paul, 1985; Siegel, 1988) can fit into the framework. For example, they can supply discrete traits and more logic-based operations to the information processing and criteria elements, and can enhance the elements through metacognition, or by expecting the critical reflection element to be applied after each of the stages, to identify any flaws in the thinking that has just been done.

These are just a few examples of how existing CT theory can be aligned to the ACT Framework. It is also possible to align the ACT Framework to existing education policies on CT. For example, Table 2.4 and Table 2.5 show how the ACT Framework can be used to represent ideas about CT from Australian policy. However, what the summary in Table 2.4 and Table 2.5 do not make evident is that the Australian Curriculum, Assessment and Reporting Authority (ACARA) has a stated position on critical and creative thinking which they have included in the national curriculum (ACARA, 2013), whereas the tertiary education sector does not (Jones et al., 2011). In addition, the Threshold Learning Outcomes (TLO) for university education have not been specifically designed around CT. That being said, students and graduates would need to employ CT skills to develop a coherent understanding of science (TLO 1) and also to achieve proficiency in the statements covered in the "inquiring and problem solving" (TLO 3). The CT expectations by the Office for Learning and Teaching (OLT, formerly the ALTC) might be implied rather than stated. Another difference between ACARA and the OLT policies is that the TLO's do not place the same emphasis on reflective skills. However, university graduates are expected to be able to critically analyse and solve scientific problems. This requires students to analyse and evaluate the outcome when they:

- search for scientific information (TLO 3.1);
- design experiments (TLO 3.2);
- interpret and draw conclusions from scientific data (TLO 3.3).

The ACT Framework approach and definition is powerful because it allows for multiple combinations of existing ideas on CT to be incorporated within it. Essentially, the limits are the understanding of the framework user. However, the ACT Framework is also a tool that can help facilitate increased understanding about CT, because as well as outlining the CT process, it supplies core terminology and concepts about CT which can be used to form questions and provides search criteria for exploring the literature. During the consultation process for the Foundation -10 Australian Curriculum, respondents noted the need for more explicit descriptions and examples for the general capability critical and creative thinking (ACARA, 2011). Respondents also found some of the terminology too specialist and wanted a glossary. This further emphasises the need for a tool such as the ACT Framework to provide terminology and understanding about CT for Australian educators.

Table 2.4 ACARA's ideas about CT aligned to the ACT Framework elements. *Note*. Information and wording of curriculum ideas extracted from ACARA (2013).

Framework Element	Related Australian curriculum ideas
Purposeful querying	• "Students pose questions" (Inquiring organising element); and • "consider and expand on known actions and ideas" (Generating ideas organising element)
Using information processing skills	Expected thinking skills would include: • "inquiry skills"; "investigate and analyse ideas and issues"; "collect, compare and evaluate information" (Inquiring organising element); • "explore situations"; "assess options"; "generate" and "consider alternatives"; "seek solutions" (Generating ideas organising element); and • "[use] logic and reasoning"; "evaluate" "draw conclusions" and "differentiate components of decisions made and actions taken" (Analysing organising element).
Applying criteria	Discipline areas may refine these, but broadly, the expectations are: • "[explore] information from a range of sources" (Inquiring organising element); • "assess options and actions when seeking solutions" (Generating ideas organising element); • "reflect on actions and processes"; "identify" and "explain thinking"; "apply knowledge gained in one context to clarify another" (Reflecting organising element); and • "apply logic and reasoning"; and "evaluate procedures and outcomes" and "assess ideas, methods and outcomes against criteria" (Analysing organising element).
Using existing understanding	The previous content knowledge, understanding and learning that students bring with them.
Critical reflection	ACARA's stance on this involves "students reflecting on, adjusting and explaining their thinking and identifying the thinking behind choices, strategies and actions taken… [through] think[ing] about thinking (metacognition), reflect[ing] on actions and processes, and transfer[ing] knowledge into new contexts to create alternatives or open up possibilities" (Reflecting organising element).

Table 2.5 Alignment of OLT's ideas about CT from the Threshold Learning Outcomes (TLO's) to the ACT Framework elements. ***Note*****. Information and some wording about the TLO's extracted from Jones et al. (2011).**

Framework Element	Related Science TLO ideas
Purposeful querying	Science graduates are expected to be able to: • Demonstrate a coherent understanding of science (TLO 1) • Exhibit depth and breadth of scientific knowledge (TLO 2) • Critically analyse and solve scientific problems (TLO 3) Therefore, 'querying' should include exploring information for the purpose of developing and increasing understanding of scientific knowledge, scientific methods and scientific rigour.
Using information processing skills	Expected thinking skills would include: • "gathering, synthesising and evaluating" (TLO 3.1) • "designing and planning" (TLO 3.2) • "selecting and applying practical and/or theoretical techniques or tools" (TLO 3.3) • "collecting, accurately recording, interpreting and drawing conclusions" (TLO 3.4)
Applying criteria	Discipline areas may further refine these, but broadly, the expectations are that graduates can: • explain why science is contestable and testable by further inquiry (TLO 1.1) • explain the role and relevance of science in society (TLO 1.2) • "select" and "apply" "practical and/or theoretical techniques" (TLO 3.3) This means students should draw on criteria from the scientific methods and theoretical standards and "regulatory frameworks" (TLO 5.3) to solve scientific problems. Students are also expected to "evaluate information from a range of sources" (TLO 3.1).
Using existing understanding	Graduates are expected to • Have "well-developed knowledge in at least one disciplinary area" (TLO 2.1) • Have "knowledge in at least one other disciplinary area" (TLO 2.2) • Be able to "explain why science is contestable and testable by further inquiry" (TLO 1.1) • Be able to "explain the role and relevance of science in society" (TLO 1.2) Students are expected to have and bring scientific understanding with them as they undertake CT to challenge new information.
Critical reflection	Reflection and metacognition are not mentioned in the TLO's.

2.3.6 Comparison of the ACT Framework definition and existing CT perspectives

The ACT Framework definition is most similar to the Facione perspective (1990a, 1990b). However, this new iteration improves on the Facione perspective in a number of ways. For example, the Facione definition is worded awkwardly. The Facione definition describes CT as a product ("purposeful, self-regulatory judgment") that "**results in**" cognitive skills and "evidential, conceptual, methodological, criteriological, or contextual considerations" about that judgment product. It is convoluted and confusing about exactly what directs what. This is frustrating because the core aspects it mentions (such as the purpose, product (judgment) and process (cognitive skills and self-regulation)) are sound and the most widely agreed upon components of CT. Consequently, my new definition includes these components but emphasises that the judgment (product) **results from** purposeful thinking and reflecting/reflection on issues related to the judgment.

The ACT Framework definition also removes two things from the Facione definition. Firstly, it strips out specific references of the core cognitive skills that were included in the Facione version; because of the temptation to focus on those aspects of the CT process at the expense of others components of the definition (such as occurred with Bloom's taxonomy, see Anderson et al., 1994). Secondly, the framework definition removes the "self-regulatory" aspect of the judgment from the core definition and replaces it with reasoned, reflective judgments. This change was made because there are situations where the core steps to CT are applied, yet the judgment was made without self-consciously monitoring, self-examining (in the metacognitive sense) or self-correcting. For example, school or university assignments may ask the student to 'critically evaluate' or 'explain the implications and consequences of X (insert course related to discipline area).' In these instances, the student will deploy CT skills and criteria in response to the set question. They have not had to generate the initial query, and in focusing on the assignment, may not think about analysing and evaluating "one's own inferential judgments" and "cognitive activities" (Facione, 1990b, p.19) in relation to the task, because they are thinking about the task itself. Their goals relating to thinking performance are focused on meeting assessment criteria, rather than seeking to be self-corrective. Further, unless they are allowed to resubmit an assignment (or submit drafts before final submission), they may not take the time to reflect on "deficiencies" and "remedy" them (Facione, 1990b, p.19). None of this is in the spirit the Facione definition ascribes to the self-regulatory aspects of CT. With regard to self-examination, Facione construed that this involved thinking about one's own mindset and interpretations and thought construction around an issue (1990b). Whereas the reflective thinking undertaken as part of CT could also involve thinking about the construction of someone else's ideas about an issue - which is reflective, just not self-reflective. While self-regulation is an important trait for being a critical thinker, it is not necessarily needed for doing CT, because it is possible to do CT (in the sense of being reflective and examine reasoning and motivations around an issue) in ways other than Facione described it.

Overall, by simplifying the language and removing details (but not the core aspects) of the Facione perspective, this new definition seeks to clarify what the CT process involves; what the CT process produces; and why the CT process is undertaken in a way that previous definitions have not described. The ACT Definition and Framework are not intended to be used in exclusion of existing perspectives. The ACT Framework sets the breadth or the minimum parameters that CT falls within, but the users set the maximum parameters (or the depth required for thinking to be truly critical in that context). For example, the details pertaining to the elements in the ACT Framework can be modified based on preferences for a particular school of thought, or cater for domain-general or discipline-specific situations, where certain thinking skills may be prioritised over another. When using the ACT Framework, educators should either: use the framework elements as a means to start exploring CT and developing their contextual understanding about CT; or they should map their existing understanding to the framework to clarify if this understanding is comprehensive, or incomplete.

How else is this new definition and framework different from others? The ACT Framework Definition draws together existing ideas about CT to clarify the process of CT. It includes the purpose (inquiry process to resolve points of uncertainty), product (useable knowledge based on reasoned, reflective judgment) and process (deployment of thinking skills and criteria to make the reasoned, reflective judgment). It essentially frames CT as a gathering of thinking processes, knowledge manipulation, knowledge generation and reflection used to resolve the understanding gap between some existing knowledge when faced with ‘new’ information (‘new’ being relative to the individual undertaking the thinking). Many of the previous definitions have had fewer aspects, emphasising some part of the cognitive, motivational and self-regulatory aspects, but not all. The ACT Framework definition itself is also unique in that it is meant to be used in conjunction with the framework elements, whereas other definitions and frameworks have been stand-alone.

The adaptability of the ACT Framework, and the accompaniment of the supporting definition, help this framework to overcome the deficits of previous models. For example, by having a supporting definition to outline the purpose of CT, the ACT Framework overcomes some of the shortcomings of previous frameworks, which either did not define CT, or did not outline how to apply the framework in relation to the CT definition (Rabu et al., 2013; Dwyer et al., 2014; Thomas and Lok, 2015). Yet despite having a definition, the ACT Framework elements are also intentionally vague to avoid limiting or isolating potential users by being too specific (or explicit) to accommodate other perspectives (Paul and Elder, 2006a; Van Gyn and Ford, 2006). Essentially the ACT Framework is highly accommodating to the existing perspectives on CT, because the non-specificity about the broader characteristics in the definition means the details pertaining to the appropriate criteria, thinking skills and judgments can be shaped by the learning domain or epistemological viewpoint of the educator or theorist.

Other limitations of previous frameworks were that some lacked certain components of the CT process, or only applied them in the context of the teacher's thinking rather than the student. The ACT Framework may initially be seen to lack important components of CT (self-regulatory components such as metacognition) because the elements are not conceptualised with the same details or emphasis that some of the perspectives place on them. However, unlike other frameworks, the ACT Framework also explains why a component such as metacognition is considered to be an enhancing aspect rather than a core element (see Section 2.3.1). Further, because of the adaptability of the ACT Framework, users can emphasise the role of the elements in line with their conceptions. For example, they can emphasise that the critical reflection is highly important, and that critical reflection would involve quality checking one's own thinking for the CT to be up to the required standard in their context.

In addition, the ACT Framework helps to overcome differences between its conception of CT and the user's interpretation of CT, by enabling adjustment to the elements. Specifically, users can include additional aspects they view to be essential to CT through the 'applying criteria element.' They just need to turn the attribute into an action or a step. For example, framework users who are not convinced by the argument for the inclusion of critical reflection (instead of metacognition) can overcome this apparent lack of suitability by including metacognitive features in their judgment criteria. They might add things like 'checking thinking for…' or 'using {insert specific metacognitive skill} to improve the standard of thinking by reviewing/identifying/exploring for…' etc. to the 'applying criteria' element as a way of adding this reflective behaviour into the stages of the CT process.

The final advantage of the ACT Framework (and supporting definition) over other frameworks, is that the ACT Framework highlights adaptability as a core attribute (Section 2.3); whereas other creators failed to explain (or realise) how their framework could be adapted (Beghetto, 2002; Edwards, 2007). This means that users of the ACT Framework should be clear about what they can and cannot change to make the approach suit their unique context. For example, in this framework, what constitutes 'appropriate' in terms of the skills and criteria can be shaped by the learning domain or epistemological viewpoint of the educator or theorist. However, the purpose of CT (generating understanding through reasoned, reflective judgments) and the broader characteristics (having criteria, being reflective, and doing some information processing) have to remain as these distinguish CT from thinking, reflecting or judging.

The ACT Framework and its supporting definition are designed to help users gather the various perspectives on CT and make this knowledge both accessible and usable. However, the framework has broader applications than providing a way to categorise and organise existing ideas on CT, because by reflecting on the core components of CT and the process of CT, it opens up possibilities to examine one's own (and other's) thinking and identify opportunities for intervention.

2.4 Value of ACT Framework for educators

The ACT Framework can help educators to clarify their view of CT. It provides terminology to help users sort and seek out ideas about CT, making CT a more accessible concept. The framework elements also accommodate users existing understanding of CT, so educators can gauge how complete their existing understanding of CT is and readily identify any gaps of weaknesses in their conception.

The ACT Framework and its elements help explain the CT process. It emphasises the treatments that a thinker applies to information to develop a deep understanding of the issue and/or make reasoned judgments about it. Educators can take advantage of this structured explanation to assist with curriculum design and evaluation. For example, with curriculum design the framework may shape how knowledge is delivered – what prompts are given or withheld to initiate the purposeful querying; how knowledge is treated - can the knowledge be challenged;

and define the expectations about how to challenge knowledge and make sound judgments within the bounds of a particular discipline.

The ACT Framework assists educators to make decisions about how to present CT within their classroom. For example, when setting a learning task or assignment, the requirements for purposeful querying and applying criteria elements could be explicitly stated or embedded within a task. Alternatively, the decisions could be handed to the students, to allow them to identify the problem and/or determine the steps they need to take to resolve it using previously learned skills. In terms of Ennis's views on instructional approaches (1989), the general approach in the framework would involve explicit descriptions of generic criteria to apply to the information during the inquiry process, and any discussion about the overall inquiry process would focus on the nature and components of CT. An infusion approach would include the provision of CT terminology, in addition to relevant discipline-specific criteria and information processing skills. An immersion approach would be similar to the infusion approach, but would exclude the CT terminology (as generic CT principles are not made explicit in this pedagogy). Whereas the elements, when shaped by a mixed approach (a combination of domain-general teaching and content-embedded teaching) would depend on the focus of the teaching for that specific lesson. A mixed approach would involve examples from both course content and real life, where the educator is explicit about both the domain-general and discipline-specific CT criteria and information processing skills needed for CT. In all of Ennis's classifications of instructional interventions, the purposeful querying and critical reflection components have the potential to be either self-directed or context-led, with the language and content used to intentionally initiate these steps depending on whether CT is to be treated independently or embedded within other disciplines.

Concerning curriculum evaluation, the ACT Framework can be used in part or full to help the educator understand students' CT development and ability. For example, assignments can be structured to step the students through some of the framework treatments to reveal the student's ability with that aspect of CT. Educators might reflect on whether their students know what the element involves, and/or know how to apply it. If educators structure an assignment that involves all the elements in the CT process, then they can both evaluate the student's understanding of the course content and principles, as well as their CT ability. They can start with questions like - Does the student make it through all the CT stages? Is the understanding produced at a high standard that is appropriate to context? However, if educators are more interested in finding out how a student's CT abilities might apply outside the classroom context, then they could consider deploying a CT test - one that takes the students through all the framework stages and reveals their abilities without the familiar content.

The ACT Framework is not simply useful for educators because it gives them essential CT terminology, but it is also valuable to them because it can be structured to help guide their pedagogy, including delivery and assessment strategies. Further, because the focus of the framework definition is about CT transforming information into useable knowledge, teachers should be able to see how this approach enables them to design a curriculum that helps to develop students' CT and meet their content goals (rather than developing CT at the expense of teaching other content).

The benefits of the framework are not limited to helping educators. It will also enable their students. In 2013, Moore stressed the importance of dealing with "institutional meta-languages" and terminology that "are fundamentally 'polysemous' in nature" to help students identify where CT fits "into their work" (Moore 2013, p.519-520). CT can mean different things in different contexts, and the various versions of the activation of the framework elements can help highlight these differences while maintaining common terminology. Observing these various expressions of CT could help give students an appreciation of the differences. However, if the educator is also aware of where their own perspective sits in the framework activation examples, then they can enunciate their preferences to the students. Therefore, the benefits of the simplification of language in the framework, as well as being able to use the framework to increase awareness of the various stages of thinking an individual requires to go through to arrive at well-considered judgments, will aid students in their CT development.

2.5 ACT Framework as relevant to my thesis

The studies undertaken for this dissertation included exploring both educator perceptions and educator actions in science in Australia. Therefore, science is the context framing the expectations around criteria, critical reflection and information processing skills. The 'information processing skills' relevant to the application of the ACT Framework are consistent with the thinking skills and dispositions valued by the Australian Learning and Teaching Council (ALTC) and Australian Curriculum, Assessment and Reporting Authority (ACARA). These skills include planning, posing questions, sound decision-making, independent problem-solving, inquiring, assessing validity, integrating and interpreting information, reasoning, reflecting and evaluating, constructing defensible evidence-based arguments, and information adaptability (Jones et al., 2011; ACARA, 2013). These are overarching policies currently in place for the Australian primary/secondary (ACARA) and tertiary (ALTC) sectors. Therefore, working from these views facilitates meaningful investigations for science educators in Australia. However, by using the policy perspectives as a baseline, there is an underlying assumption that Australian science educators share these policy views. To verify this, study 1 (Chapter 3) explores educator perspectives on CT to see what the current level of understanding about CT is, and which aspects of the policies aligned with practitioner views. This is complemented by the two case studies that form study 2 (Chapter 4, 5 and 6), where specific educator actions are investigated to measure the effects that current approaches in science education have on students' CT abilities. The results provide awareness about what Australian science educators know about CT and what they are doing to help develop students' CT skills. Then the application of the framework to the findings from these studies provides a model for educators to follow, so educators can employ this approach to identify areas in their own practice that need improvement.

2.6 Conclusions

This chapter is the second of three chapters that explore the nature and conception of CT. It has presented an original reconfiguration of the literature on CT and resulted in the synthesis of a new definition and framework. This framework is unique because it is adaptable to context, as well as being able to be used as a stand-alone definition for CT. Through the next three chapters, the power of this framework to further inform classroom-based understanding of the nature of CT and its development is demonstrated.

PRIMER 2

Transdisciplinary research borrows knowledge from one discipline and moves it into another knowledge system, with reflexivity and consciousness. The imperative is to challenge, transform and build new insights, paradigms and approaches to a topic, problem or issue. The previous two chapters explored critical thinking (CT) theory through the eyes of a scientist trying to understand a new discovery. With CT theory spanning multiple fields, it was necessary to look for common aspects, as well as differences, to gather evidence to formulate a perspective to unify the ideas. I could see that the emphasis on different aspects of CT varied across the fields, but there were aspects to the process which prevailed irrespective of discipline. CT is like describing the colour of an object, both have subjective nuances even though it may be described with the same word.

Aside from the thinking required to synthesise and organise the elements, much thought went into the display of the resulting framework. At first, I drew on my knowledge of shapes. Could it be represented as a funnel, cogs or a formula? All of these inferred connections but the form of many of these shapes implied linearity, which CT is not. The idea of a 3D prism shape lingered for a while, but the inferences resulting from the edges of each face became troublesome. I started to question what connections and interactions they implied and found I lacked summative terminology to explain them. For example, there is not a word to describe thinking that occurs when concurrently applying criteria and using existing knowledge, especially one that can also communicate that the criteria and the knowledge can change, and that even though someone is not consciously using other framework elements (such as critical reflection, purposeful querying, using information processing skills) they are still there ready to be activated at a moment's notice. All of these things were also hard to communicate visually through the face or edge of a 3D shape. Next, I started thinking about lightbulbs and switches, and after a conversation with one of my supervisors about sheet music and math formula's I decided the simplest visual display to convey both interactions, change and dormancy was a display of elements over multiple lines, with colours to infer element activation at each line. Thus, the visual construction of the framework was inspired by knowledge from multiple fields.

This next step in this transdisciplinary journey sees the incorporation of social science research methods (a perception survey with Likert items) with some CT theory and education policy (survey item themes), which were then sent out to stakeholders (Australian educators). The responses were then brought together with some biological analysis techniques (in the form of multivariate analysis using Primer software). This mix of methods enabled the generation of some preliminary educator profiles that characterise the way that different groups of Australian educators think about CT. This is the first time such an approach has been taken, and this understanding would not exist without transdisciplinary thinking.

Chapter 3

Educator Perceptions of Critical Thinking

We shall make surer and faster progress when we devote ourselves to finding out just what education is and what conditions have to be satisfied in order that education may be a reality.

Dewey, 1986, p.252

3.1 Chapter outline

This chapter is the third in a series of explorations into the nature of CT. This chapter shifts this exploration of CT theory into the classroom – a perspective largely missing from the literature. It presents a study aimed to establish an Australian perspective about science educators' current level of understanding of critical thinking (CT) and its development. The chapter explores educator responses to a specifically designed survey about the nature of CT, the development of CT and the assessment of CT. It also explores responses about the nature of CT using the Adaptive Critical Thinking (ACT) Framework described in Chapter 2, to demonstrate how this theoretical contribution helps clarify understanding about CT. This chapter contributes answers to Research Question 1: *What ideas are represented in Australian educators' perceptions of critical thinking and what approaches do educators report incorporating in their classrooms to develop and assess CT?* This study also sought to answer some more specific questions about the difference in educator viewpoints by discipline (science versus non-science) and education setting (high school versus tertiary), to determine how an educator's background might influence the ideas they hold and the choices they make. These questions were presented in the introduction, but restated here for convenience:

a) What do Australian science educators know about the nature of CT, and is their understanding different depending on their discipline or education setting?
b) What approaches do Australian educators report incorporating in their classrooms to develop CT, and are these approaches different depending on discipline or education setting?
c) What methods and tools would Australian educators use to assess CT, and are these different depending on discipline or education setting?
d) Do factors such as teaching experience, educator location (as a proxy for identifying state-based policies), or gender, help explain differences in educator perceptions about CT?

The first part of this chapter explains the context, design and distribution of the survey; the second half explains the analysis approach, the results and discusses the key findings and implications. Through identifying and exploring the views educators held about CT, it was possible to determine that discipline and education setting help shape educator perspectives about CT. However, this study also found that educator characteristics such as discipline, education setting, state, gender and teaching experience were not enough to explain all the differences in educator perspectives. As such, a multivariate analysis approach was used to explore for natural groupings that classified educators into clusters based on their perspectives, irrespective of their demography.

3.2 Introduction and educator perception study research context

The inclusion of CT in the Australian curriculum and university Threshold Learning Outcomes has increased educator's responsibility to develop CT skills in Australian students. However, these policy documents fail to supply a definition of CT to guide educator conceptions. CT is a multifaceted construct with widely debated attributes. In addition, numerous meta-analyses have revealed a range of approaches is used for developing and assessing CT (Allen, Berkowitz, Hunt, and Louden, 1999; Ortiz, 2007; Niu et al., 2013; Abrami et al., 2015; Huber and Kuncel, 2016). Therefore, the sheer number of perspectives and approaches, in addition to the complex nature of CT, can make it difficult for educators to decide which theories and strategies to integrate into their pedagogy and practice.

Numerous studies have explored how educators perceive and develop CT, and have identified uncertainty and low confidence regarding teaching for CT (Shell, 2001; Choy and Cheah, 2009; Black, 2009; Kowalczyk et al., 2012; Aliakbari and Sadeghdaghighi, 2013). For example, Black (2009) found that UK educators struggled with the introduction of mandatory CT curricula due to timetabling issues (relating to teaching loads and also teaching time), motivation and resistance from students, as well as a lack of support (i.e. resources and training). Lack of knowledge and tools is a familiar theme in CT literature (Lauer, 2005), but knowing how to provide opportunities to develop CT is only part of the problem. Lack of knowledge also includes a lack of understanding about the nature of CT, because there are differences in the emphasis of particular components of CT within the various theories about CT. Stedman and Adam's study (2012) found that there are differences in educator understanding of the nature of CT. In this study, all of the life science educators (n=51) indicated they thought CT involved higher-order thinking skills, but only 78% of those same educators thought it included a reflective thinking component. However, Stedman and Adams failed to note that these differences are a product of the various perceptions that exist in the literature. It is not just enough to be aware of differences, there is also need to identify where these differences come from, to better support educators to develop CT in a way that is relevant to both their existing understanding and their educational context.

There have been three published studies addressing educator's CT perceptions conducted in Australia (see Jones, 2006; Moore, 2013; Danczak, Thompson and Overton, 2017). The first Australian study considered five disciplines (economics, history, law, physics, and medicine) across two institutions (Jones, 2006, 2009, 2013); though some of the papers only compare some of these disciplines (such as Jones, 2007). This study reports that way an educator conceptualises "the epistemology of their discipline" is core to shaping their "understanding and teaching of generic skills" including CT (2006, p.220). The second study focused solely on conceptions about CT by history, philosophy and literary/cultural studies educators from a single university (Moore, 2013). This study found that educators favoured a multidimensional view of CT. However, Moore also found that there were differences in perspectives about certain aspects of CT meaning even within a sub-set of related humanities disciplines. The last and most recent Australian study considered the views of tertiary chemistry educators, students and employers (Danczak et al., 2017). This study found differences in student, educator and employer understanding of CT concerning skills that are developed through and/ or desired outcomes of studying chemistry. However, there are still many unknowns when it comes to how Australian science educators perceive CT and how CT abilities develop in their classrooms. Therefore, this first investigation aimed to explore Australian high school and tertiary science educator's ideas about the meaning of CT, developing CT and assessing CT.

There are differences between high school and tertiary education environments. For example, they are governed by different policies. In Australia, high school educators have both mandated curriculum and school-based policies, whereas tertiary educators have more freedom and control over their course content and course delivery. Similarly, there are disciplinary-based differences between physical sciences compared to humanities and arts topics. With the first relying on numbers and objectivity for evidence, while the other gathers evidence through words, imagery and contextual interpretation, therefore having more subjectivity. Yet, despite the conventional wisdom that there would be differences in perspectives by demographic factors such as discipline and education setting, given that the very nature of different disciplines and different education setting mean there are different criteria which shape decision making. However, this has not been confirmed statistically. Therefore, the rationale around the educator background choices are as follows: -

To investigate if education settings and/or disciplines influenced philosophies of CT, potential participants were identified based on the educator's teaching setting (high school versus tertiary institutions) and discipline area (science versus non-science educators). These variables were of interest to this doctoral research because educators from high schools work from a national curriculum and therefore, may be similar in their approaches. Conversely, educators from universities work with different and often self-set curriculum. Therefore, it was expected that there might be some differences in the views they hold.

With regard to the science versus non-science educator comparison, the non-science educator discipline (history) was selected to serve as an outgroup, to evaluate whether the approach taken by science educators was vastly different from non-science educators. History has a comparable course offering between secondary and tertiary education settings, which made it an ideal choice for this study. History, like science, has no set definition for CT. Previous studies have explored the views of history educators (Moore, 2013) and science educators (Danczak et al., 2017), but have not explored potential differences between the views of science and history educators. Theorists often debate about CT's discipline-specific nature (or lack thereof), so since these two disciplinary groupings have different methods and rules for gathering and interpreting evidence, it was thought it would be interesting to see if this translated into a different view of CT.

While educators were targeted by discipline and education setting on a national level, there are additional aspects of an educator's demography that could contribute to their philosophies about CT. For example, because there is no nation-wide mandate concerning CT in Australia at a senior secondary or tertiary level (except for South Australia where they are enacted - Government of South Australia, 2015), some additional attributes including state (as a proxy for identifying state-based policies) and teaching experience were considered. These factors were chosen to explore teacher-training based characteristics that might influence their ideas and incorporation of CT in their classrooms. Gender was also explored since certain teaching areas are dominated by a particular sex. Gender is also thought to play a role in survey participation (Smith, 2008).

3.3 Australian educator perception study methods

This new study was developed while recognising five key international papers (Paul, 1996; Shell, 2001; Black, 2009; Choy and Cheah, 2009; Stedman and Adams, 2012). Each of these international papers reveal different aspects of educator views towards CT and its development. In brief, Choy and Cheah (2009) published a qualitative study of 30 Malaysian tertiary educators from various disciplines, consisting of eight items about educator perceptions of CT. The work of

Stedman and Adams (2012), adapted the qualitative items from Choy and Cheah (2009), resulting in a quantitative study of 56 self-selected tertiary agriculture and life science educators from one USA institute. The authors asked educators about CT concepts and instruction through 21 assessable items, consisting of Likert-type scales and multiple-choice questions. Black (2009) produced a more extensive study on the challenges of introducing a CT curriculum in high schools. Her study employed a mix of 50 qualitative and quantitative items to explore the views of 236 UK high school educators who teach CT courses. Lastly, Shell (2001) quantitatively explored the views of 176 tertiary nursing educators teaching in generic baccalaureate programs in the USA. This study used the most items, comprising of 52 assessable items, 43 of which were Likert-type scale questions. This study focused on barriers to CT development, which is a useful means for identifying where limitations exist and resources needed. However, Paul (1996) noted that educators do not always grasp that they have gaps in their understanding of CT. Asking questions purely from a barrier perspective could miss opportunities to identify bigger issues related to educator understanding about the construct of CT, which would also influence how they approach and incorporate CT in their pedagogy.

Few Australian studies have explored educator conceptions (see Jones, 2006; Moore, 2013; Danczak et al., 2017). However, these studies all used a qualitative methodology, with open-ended survey questions and/or interviews. Also, none of these Australian studies considered the perspectives of senior secondary educators. Given this previous work, this current study aimed to build on the strengths and weaknesses of these papers, in order to establish a quantitative perspective about CT for Australian science educators at both a secondary and tertiary level.

3.3.1 Survey design and generation of question themes

A survey was designed to capture the beliefs of Australian educators about CT (see Appendix A: Survey Template), as well as their perceptions of the actions they take to develop and assess it. The questions were reviewed numerous times by the supervisory team and external colleagues over the course of six-months. Care was taken to balance the length of the survey and the cognitive demands of survey questions, whilst providing respondents with a mix of free-responses and scaffolding to allow them to reflect on and explain their perspectives about CT. Question styles were inspired by previous work by Shell (2001), Choy and Cheah (2009), and Stedman and Adams (2012). In these previous studies, Shell (2001) and Stedman and Adams (2012) both opted for a Likert-type scale approach, whereas Choy and Cheah (2009) used an open-ended approach. The Australian educator perception survey comprised of a mix of categorical, quantitative and qualitative question formats. The categorical/quantitative options were given to help educators frame and express their ideas about CT. This choice was made because in previous studies of educator actions teachers have indicated that they, "know it [inquiry] when they see it" (Marshall, Horton and White, 2009, p.48), but found it harder to describe without a prompt. Whereas, extended response items[2] were given to provide respondents with the opportunity to elaborate on previous responses about certain aspects of CT, and raise any elements relevant to their conceptions of CT not listed in previous questions. For example, if the respondent selected the supplied statement "I look for specific evidence of critical thinking by students in my courses," a follow-up extended response question would enable them to describe the types of explicit evidence they look for and/or the assessment tools they use.

The rationale behind the theme of the survey was also important. Several studies have previously examined barriers to CT development, and they always seem to relate to time, money, training (for example, Shell, 2001; Black, 2009). These are important factors to note when implementing interventions. They do not bring awareness of the current level of expertise that educators have. Because even if educators think they do not know much about the construct of CT itself, or recognise that they need more resources, awareness of these factors does not reveal their specific knowledge gaps, nor whether the resources they think they need will bring the best cost-benefit outcome. By way of the recruitment process in these barrier-focussed surveys (educators teaching CT), we can infer that these educators are gathering a working knowledge of CT as they continue to make choices for their classroom. Yet despite these uncertainties and limitations, educators continue to implement strategies in their classroom with the intent that they will have a positive effect on CT development. However, it is hard to know if the approaches are working because these barrier surveys fail to capture details about the strategies. To overcome this limitation, the content of the survey in this new study seeks for the level of educator understanding about CT so suitable interventions and training can be developed rather than looking for reasons why teachers are finding it hard. In particular, this new study contributes to the existing body of knowledge by increasing understanding of science and history educator's knowledge and practice concerning CT.

[2] The extended response items are essentially open-ended response items. However, I have specifically described them as extended response items because their purpose was to inform other responses, rather than be standalone responses.

The themes incorporated in the Australian perception survey were a blend of literature and policy-based ideas about CT conceptions, CT instruction and CT assessment. As with the question styles, some themes were also inspired by Shell (2001), Choy and Cheah (2009), and Stedman and Adams (2012), as well as from Australian policy. For example, Shell's (2001) study addressed "barriers that educators may face when implementing critical thinking teaching strategies" (Shell, 2001, p.288). The final version of the survey did not include any of Shell's wording, but it did incorporate some of the themes raised in Shell's study regarding the difficulty of recognising CT and difficulty of assessing CT. The inclusion of questions of this nature helped to anticipate some of the challenges Australian educators might face when trying to develop CT skills in their students.

Choy and Cheah's (2009) study explored educator conceptions of CT, educator perceptions about the ability of students to think critically, and also perceptions about the role of educators when incorporating CT in the classroom. The findings of their study indicated that educators had misconceptions about evidence of CT abilities in students. That is, not just a matter of comprehension. Further, educators faced a conflict when making decisions about completing the stipulated course requirements and being able to develop student CT abilities. Thus, modified versions of Choy and Cheah's qualitative questions were included to provide a more detailed understanding of educators' perception and practice.

Stedman and Adams (2012) instrument adapted the questions used in Choy and Cheah's (2009) study. They supplied 15 statements that asked educators to indicate the extent of their agreement (or disagreement) about CT and CT instruction (Stedman and Adams, 2012). Based on Stedman and Adams findings, it was anticipated that responses to these statements would reveal Australian educator beliefs about the skills associated with CT, and their understanding of their responsibilities in developing CT skills. Further, since Stedman and Adams study (2012) included some life-science academics, using a similar question style meant some findings of this study were readily comparable to views held by other science educators.

Additional content was included in the survey to incorporate some of the specific CT skills outlined in the Australian policy documents, such as inquiring, inferring, interpreting, thinking broadly and deeply, assessing evidence, drawing conclusions and problem-solving. This Australian policy content was incorporated through question relating to the nature of CT and the development of CT.

The final survey (see Appendix A) comprised of six core quantitative question sets in three main question formats: two Likert-type scale question sets (Questions 2 and 4), two point-spending[3] (Questions 1 and 6) and two 'select an option' item formats (Questions 3 and 5); in addition to six supporting extended response items (Questions 1a; 3a, 3b or 3c; and 6a). However, study participants received just one of the three versions of the extended response follow-ups to Question 3; meaning a total of nine questions plus demographics (discipline, education setting, years of teaching, state, gender).

In terms of specific question content, statements from Questions 1, 2 and 3 were compiled to be representative of both elements from Australian education policy, as well as the core literature perspectives on the nature and development CT. Question 1a (qualitative) allowed respondents to express their understanding of the nature of CT by listing features, skills, traits and dispositions they associated with it. The qualitative follow up to Question 3 had three different versions depending on the development approach respondents selected as their answer to Question 3 (see Appendix A: Survey Template: Questions 3 and 3a-c). All the qualitative options asked for specific examples and explanations about the respondent's practice, but had phrasing relevant to the option the respondent had chosen in Question 3 (either explicit instruction, embedded approaches, or no intentional actions). Question 4 listed several teaching strategies approaches associated with CT development, as well as emerging teaching innovations (i.e. flipped classroom), allowing respondents to think about their own experiences and what they have found to be effective with their students. Question 4 also included the option 'other,' which allowed respondents to specify any other strategies they have used and if the strategy was effective or not. Questions 5 and 6 listed strategies and tools that are used in the assessment of CT, allowing respondents to reflect on their own choices and strategies when it comes to assessment. Question 6a allowed respondents to explain any additional assessment strategies they would use. Question 3a also asked about assessment. However, it was only given to a subset of respondents (those who selected explicit instruction), to increase understanding of their overall intentionality and confidence about the effectiveness of their explicit assessment approach. Questions 5 also

[3] A 'point spending question' refers to a question format where educators were given points to assign across a set of statements based on how close each statement was to their viewpoint. The survey instructed that statements which were not relevant should be assigned a zero, and the number of points assigned to a relevant statement should indicate how strongly that statement fits their view on that aspect of CT (see Questions 1 and 6 in Appendix A: Survey Template).

included ‘other' as a tick the box item (with an empty text box), to enable educators to add their ideas should something about their view on that aspect of CT be missing from the existing options.

3.3.2 Survey distribution

Survey responses were collected online through Survey Monkey®. A link to the survey (as well as accompanying participant information) was sent out to potential respondents via post or email. This study requested the participation of Australian science and history educators who teach in senior secondary and tertiary courses. Explanations for these target groups were outlined in Section 3.2.

Schools and universities were identified through two online directories: www.australianschoolsdirectory.com.au/ and www.australianuniversities.com.au/list/. Contact details for schools and universities were then gathered from their corresponding institution websites. The survey was posted to a random selection of 1500 schools, and follow-up emails were sent to encourage participation. The randomisation process consisted of creating a column of data using Microsoft Excel's random number generator function, and then ordering the school names by the random number they had been allocated. Permission was sought from principals prior to senior secondary educator participation, and acknowledgement of this permission was demonstrated through educators completing the survey. Survey distribution to approximately 1000 tertiary level participants was achieved through an email request to teaching and learning leaders in science and history, as well as through direct emails to academic staff listed as actively teaching into the relevant subject areas on their university website.

Participation in the study was voluntary, and responses were anonymous (except for respondents who supplied their email address so they could be contacted about results/training). Email addresses were removed from the data set prior to analysis, ensuring anonymity. The survey was approved for use by the Social and Behavioural Research Ethics Committee (SBREC) at Flinders University (Project # 6431).

3.3.3 Educator perception study analysis

Data processing

Educators supplied 179 responses. Survey responses were checked for completeness (noting how many questions and items were skipped) and for relevance to study (are respondents from a targeted discipline). Responses with one or more skipped quantitative questions or an irrelevant discipline were excluded from the data set. Responses with attempts at all quantitative questions but a few skipped items were included in the data set. These skipped items were treated as missing data and were excluded on a case-wise basis from univariate analyses, i.e. from frequency analysis. However, in the multivariate space where questions were analysed using their collective set of items, these missing items were assigned a zero. Assigning them a zero served to distinguish these respondents from those that answered that item, but not so much so that it would skew the results. This only affected a handful of participants, and in the scheme of patterns from the set of 80 items, would have had little effect on the shape of the data cloud and hypothesis testing results. After the completion of this data processing, there were 146 educators left in the study sample.

Some additional processing was undertaken for Question 5, as respondents did not seem to distinguish between published and purchased resources. This is probably due to the descriptors that did not indicate the difference between published (meaning peer-review published research) and purchased (educational resources which may not have a peer-reviewed evidence base). To correct for this, the published and purchase resource responses were recoded into 'external resources.' Lastly, to better highlight trends within Question 1 and 6 mean proportion of point calculations excluded zeros. Values were calculated to represent the average weighting of points to this statement from educators who assigned one or more points.

Quantitative analysis

As an exploratory study, the first set of analyses consisted of generating frequencies for each of the quantitative questions for the demographic factors. Next, to identify trends within the responses to the CT survey, multivariate data analysis was used to quantify and elucidate patterns across respondents. Eighty quantitative response items (herein referred to as 'items') were collated from the six quantitative response questions. These 80 items collectively formed a global profile comprised of three main theoretical elements: CT development (36 items – Questions 2, 3 and 4), CT assessment (33 items- Questions 5 and 6), and concept of CT (11 items – Question 1). Due to the variation in question design (point spending, Likert-type scale, and select an option), each item was scaled to 0-5, where the higher the score, the stronger the perspective pertaining to that item's CT theme. This scaling provided equal weight to each item in the global profile, irrespective of questioning type. Descriptions of quantitative questions and relevant item codes can be found in Appendix B: Tables B.1-B.4.

The experimental design was set-up as a two-way crossed design with demographic information (discipline, education setting, gender, state and teaching experience) being the main *a priori* factors used to explore respondent's perspectives, with the possibility to seek for various interaction effects within all possible crosses of these factors. Responses were investigated in several ways. Sample-similarity matrices using the Bray-Curtis similarity algorithm (Bray and Curtis, 1957) were generated for each of the question-sets 1-6; and for the question-sets separated into their corresponding theoretical aspects aka sub-profiles i.e. CT development (36 items: Question 2, Question 3 and Question 4), CT assessment (33 items; Question 5 and Question 6) and CT conceptions (11 items; Question 1); as well as for the global profile (80 items). Then for each of these data matrices, significant differences between *a priori* groups of respondents (discipline, education setting, gender, state and teaching experience) were evaluated using the two-way Permutational Multivariate Analysis of Variance (PERMANOVA). The data was further tested for any interaction effects using a two-way crossed design. An interaction effect seeks for consistency across multiple levels of a factor, for example, whether an observed pattern between educators by their teaching discipline is sustained within gender levels. So here, each factor was crossed with each other factor to better explore for differences between groups. Overall, this approached allowed for; exploration of interactions through a type III (partial) sums of squares approach, fixed effects sum to zero for mixed terms, and unrestricted permutation of raw data to generate exact p-values (Anderson, 2001). Groups of respondents were considered significantly different from one another if the p-value falls <0.01. This is a conservative choice to mediate for variation that might be attributable to the small sample size of the study. It was also used in place of a correction for multiple testing, to help to reduce the likelihood of a false-positive given the number of analyses conducted.[4]

When PEMANOVA analysis revealed significant differences between target groups ($p<0.01$), the Similarity Percent (SIMPER) algorithm was used to reveal the contribution that each item made towards the observed difference. Due to the number of items in the global, development and assessment sub-profiles, the overall contribution of each question-set to group dissimilarities was calculated from the SIMPER results and reported with the minimum and maximum percentage contribution of items from that set. For explaining group differences within individual questions, the top contributing items to dissimilarities between groups for key educator

[4] This is despite the fact that each of the analysis were planned and hypothesis-driven rather than undertaken in a p-hacking approach, which is when a correction should be applied.

comparisons (by discipline, education setting, and within the science cohort) are reported. This was followed by descriptive analyses that consisted of calculating frequencies of educator choices relevant to each item by each demographic factor. For questions that included point allocations (Questions 1 and 6 in Appendix B: Tables B.1, B.2 and B.4), both frequencies of choices and point allocations were calculated to help reveal trends in the response choices. In addition, some calculations were weighted to moderate for differences in sample sizes across the groups. For example, due to differences in sample size between the groups, points in Question 1 and Question 6 were converted into proportions of points (rather than presenting the number of points) for ease of comparison between educators by their demography.

Next, to explore the contributions of the sub-profiles to the global profile, and also to explore patterns of behaviours within educators across sub-profiles (i.e. conceptual, developmental and assessment profiles), the independent sample-similarity matrices were correlated using Spearman rank correlation (Clarke and Warwick, 2001) via the Mantle test in PRIMER (v.6) with 9999 permutations. Then, for sub-profiles that were found to correlate, investigations using Principal Coordinates Analysis (PCoA) with vector overlays of the relevant sub-profiles were generated to determine which particular items contributed to this trend.

Lastly, because the *a priori* factor groupings were not sufficient to separate the respondents into clearly characterised groups (due to the diversity in the educator's perceptions) a SIMPROF test in PRIMER (v.6) was undertaken to seek for significant groups regardless of the educators' background. The results from the SIMPROF were then used to define the groups within the existing PCoA ordinations. Further exploration of the group's characteristics was undertaken through the unconstrained Spearmen vector overlays, to see which of the survey items moderately to highly correlated to the data cloud and draw attention to perspectives that characterise each group. SIMPER was also used to determine the five highest-scoring items contributing to each group's identity. These helped to form educator profiles that can be used to characterise an educator's approach to CT.

These multivariate analyses were performed using PRIMER (v.6) with the PERMANOVA add-on PRIMER-E, Plymouth Marine Laboratory, UK, (Clarke and Warwick, 2001). A pictorial summary of these approaches is found in Figure 3.1.

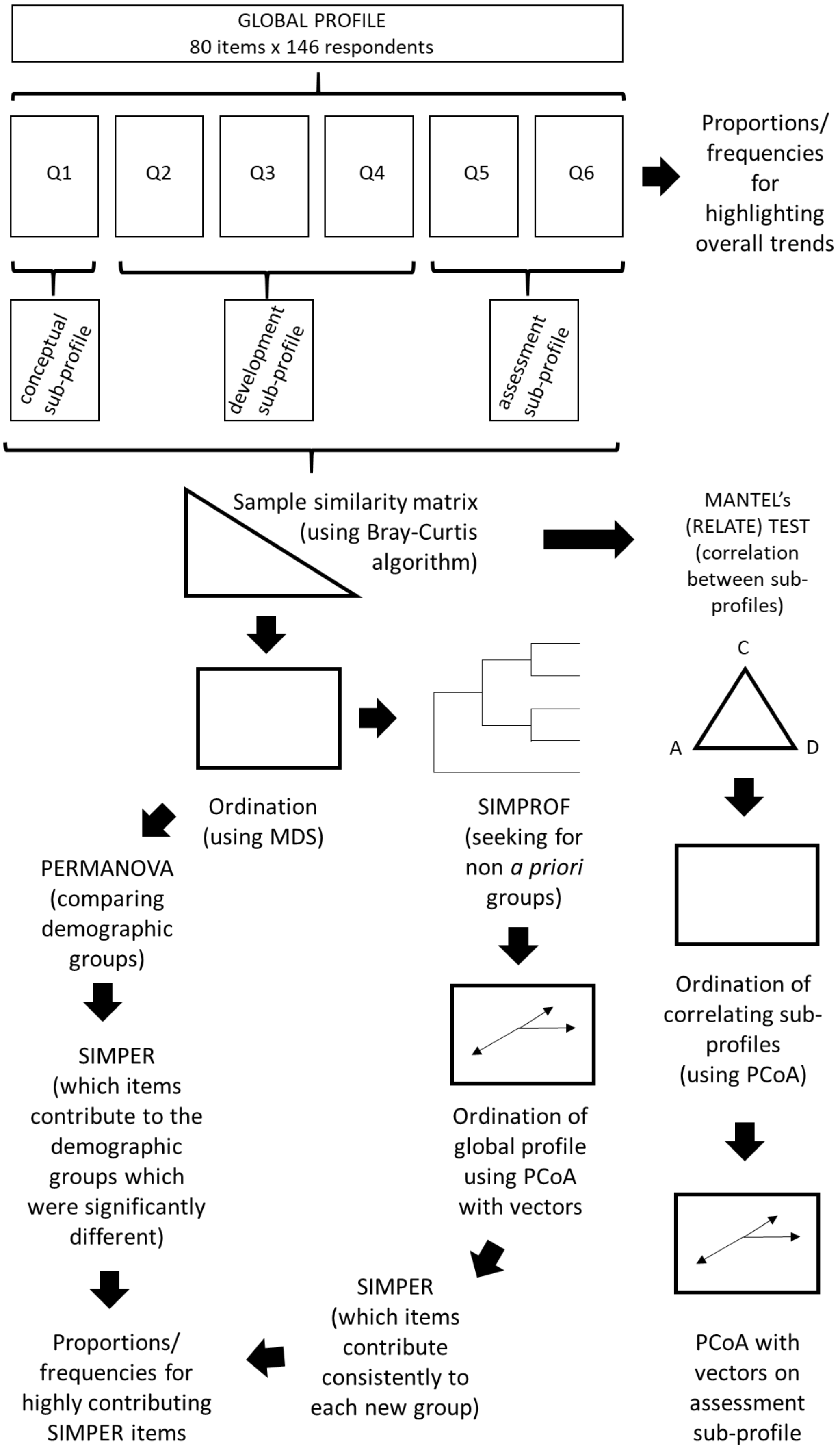


Figure 3.1 Summary of statistical approach used in the quantitative perception survey analysis. Figure created for Butler (2019).

Qualitative analysis

Questions 1a, 3a-c and 6a were designed to capture qualitative data and were expected to be coded thematically and then analysed. However, after examining responses to the study, it was evident that there was a decline in the number of responses to the qualitative items the further the participants proceeded through the survey. This meant that the sample size for Question 3a-c and 6a were too small to produce representative trends. Rather than exclude these questions altogether, the responses to qualitative questions were compiled into word clouds using Word it out. Raw responses were initially generated into a cloud using Wordle to extract word counts. Then they were copied into 'Word it out', with some additional parameters (words more than five characters, and a minimum frequency). For Question 1a the minimum frequency was set to 3. However, for all other questions, the number of respondents were much less, so the frequency was set to 1. After checking the context of words that met this criterion, words with similar meanings were grouped into themes (this helped reduce the number of singletons in the cloud). Common words were removed, and, for those words with a frequency of 1, only words which added meaning to the story about CT were included in the final cloud.

These word clouds provide insight into the repetitive themes, and therefore, some basic information for future studies. Question 1a was answered by 132 of the 146 participants, and therefore had enough responses that some additional descriptive explorations could be undertaken. Given the nature of question 1a, which asked participants to explain which feature, skills, traits and dispositions they associated with CT (see Appendix A: Question 1a), a deductive coding approach was used, with the ACT Framework and enhancing elements used as the theoretical basis. Frequencies were then generated per framework element in Microsoft Excel 2010.

3.4 Results from the Australian educator perception survey

To investigate Australian educator perceptions about CT and its development and assessment, and determine whether their understanding is consistent with Australian policy and/or the literature on CT (research objective 2), a perception survey was sent out to 2500 educators. 179 educators responded, where 65 were tertiary educators, and 114 were high school educators. However, 33 educators were from irrelevant discipline areas or did not attempt all six quantitative survey questions, leaving 146 respondents (50 tertiary and 96 high school) who completed 80 quantitative items relating to how they perceive aspects of CT (Table 3.1).

Responses were explored using multivariate analysis at a question level; as a theoretical set (referred to as 'sub-profile'); and at a global level, where profiles were constructed from the relevant questions from a participant's response set (see Figure 3.1). These profiles enabled each pair of respondents to be compared to generate similarity indexes based on their perception of that aspect of CT (using the Bray-Curtis algorithm). Then to indicate how similar respondents' overall perspectives on that aspect of CT were, a multidimensional scaling (MDS) ordination was generated to seek for patterns in the proximity between the educators. This was repeated for the ten similarity indexes (6x question; 3x theoretical set; 1x global set), with each index being explored through five demographic factors; discipline, education setting, gender, state and teaching experience (Table 3.1). Collectively, these analyses address Research Question 1: *What ideas are represented in Australian educators' perceptions of critical thinking and what approaches do educators report incorporating in their classrooms to develop and assess CT?*

Table 3.1 Summary of respondent characteristics for the Australian high-school and tertiary educator CT perception survey study.

Factors	Total	Discipline	
		Science	History
number of individuals recruited	146 (100%)	110 (75%)	36 (25%)
education setting			
- senior secondary high school (high school)	96 (66%)	69 (63%)	27 (75%)
- university (tertiary)	50 (34%)	41 (37%)	9 (25%)
gender			
- female	80 (55%)	59 (54%)	21 (58%)
- male	66 (45%)	51 (46%)	15 (42%)
teaching experience			
- 1 to 5 years' experience	23 (16%)	20 (18%)	3 (8%)
- 6 to 15 years' experience	56 (38%)	40 (36%)	16 (44%)
- 16+ years' experience	67 (46%)	50 (45%)	17 (47%)
state			
- Australia Capital Territory (ACT)	4 (3%)	3 (3%)	1 (3%)
- New South Wales (NSW)	48 (33%)	31 (28%)	17 (47%)
- Northern Territory (NT)	2 (1%)	2 (2%)	0 (0%)
- Queensland (QLD)	29 (20%)	21 (19%)	8 (22%)
- South Australia (SA)	19 (13%)	17 (15%)	2 (6%)
- Victoria (VIC)	31 (21%)	25 (23%)	6 (17%)
- Western Australia (WA)	13 (9%)	11 (10%)	2 (6%)

Since this study incorporates many different analyses, the findings are structured by themes - including the nature of CT, the development of CT, and the assessment of CT. The results section concludes with an examination of the global profile. Answers to the fourth sub-question are integrated into the other three CT areas (such as nature, development, assessment) since it concerns findings related to additional demographic factors rather than addressing a specific question about a component of CT. Overall, there were some key differences identified through two of the five measured demographic factors (education setting and discipline), highlighting that certain approaches to CT seem to have a contextual basis. However, there were other groups of respondents where the reasons driving group similarities were not necessarily attributable to any one of the measured demographic factors. This suggests there is more complexity to decision making about CT development and assessment than just context or a theoretical conception.

3.4.1 Concerning understanding about the nature of CT (Questions 1 and 1a)

There were two questions, one quantitative and one qualitative, related to science educators understanding of the nature of CT. The conceptual sub-profile was formed from the 11 statement choices in Question 1 (Appendix B: Table B.2).

Results and trends for Question 1 (aka the conceptual sub-profile)

For Question 1, educators allocated points to statements about CT to indicate which of the statements reflected their views on CT, with the number of points reflecting the strength of the alignment to their perspective. Given that this educator perception study was interested in the discipline of science, where history educators were surveyed as an outgroup, it was first important to establish whether science educators held different theoretical perspectives to other disciplines like history. The statements most frequently chosen by both disciplines included phrasing such as "higher-order thinking skills" (item C_1c), “embedded thinking skills development” (item C_1g) and “dispositions and thinking skills that develop over time” (item C_1b) (see Table 3.2). These statements also received the largest allocations of educators’ points (having up to 15 points allocated), although they most commonly received a third of every educator’s points (i.e. Table 3.2 mode = 5 for items C_1c and C_1g). The exception to this high point spending trend on these items was for C_1b, where science educators most frequently allocated a fifth of their points (mode = 3). However, the mean point allocations were more similar (3.6 versus 3.68, see Table 3.2 Item C_1b). Generally, the patterns in frequencies and proportions of point allocated across all Question 1 items were equivalent, with the statements chosen the most frequently also getting the highest proportion of points. Overall, there was no significant difference between the perspectives held by the science and history educators (Table 3.3 pseudo-F = 0.37, p-value = 0.8442), with both groups allocating their points to similar statements in similar proportions (see Figure 3.2 A-B). Despite the difference in sample size, there was consistency in the conceptual ideas reflected in all educator's statement choices, and in the strength of their agreement to those statements.

This trend was upheld when exploring the data by education setting - i.e. high school versus tertiary; and science: high school versus science: tertiary. However, there was slightly more variation between the high school science educator and tertiary science educator perspectives, as indicated by greater differences in the proportions of the population who chose statements such as C_1b and C_1f (see Table 3.2 and Figure 3.3 A-B). There was also variation in the way the high school and tertiary educators spent points in both cohorts. For example, fewer tertiary educators included C_1i in their perspectives, yet those tertiary educators who did pick it spent more points on this item (proportionally) than the high school educators who also spent points on this item. This trend is also demonstrated by the difference in the most frequent allocation of points by each education setting, where the tertiary educators spent a mode of five points, whereas the high school educators typically only spent two points on this statement (Table 3.2). However, none of these apparent trends were identified as significantly different (Table 3.3: high school versus tertiary, pseudo-F = 0.08, p-value = 0.9689; science: high school versus science: tertiary, pseudo-F = 0.37, p-value = 0.8530).

Overall, three statements were consistently chosen by the educators (see Figure 3.4 A-B), irrespective of the measured demographic factor. This was supported by the findings from the PERMANOVA where there were no significant differences identified at the 0.01 level for Question 1 (Table 3.3) when comparing perspectives by discipline, education setting or for the other measured demographic factors (i.e. age, teaching experience or state - see Research Question 1d). However, given that this is an exploratory study seeking to identify possible trends rather than confirm them, and the fact that the 0.01 significance level is quite conservative, there are a few results from the PERMANOVA at the 0.05 significance level that were further explored.

At this 0.05 level, gender seemed to exert some influence on perspectives about CT. For example, there were some significant differences identified for gender as a main effect (pseudo-F = 2.97, p-value = 0.0121), and through the pairwise comparisons undertaken for the gender- state cross (pseudo-F = 1.54, p-value = 0.0431) and the gender- teaching experience cross (pseudo-F = 2.37, p-value = 0.0122) (refer to Appendix C: Table C. 1 footnotes). When exploring the gender-state comparisons there were seven key differences at the conceptual level, some highlighting a perspective difference between educators within a state (male VIC versus female VIC; pseudo-F = 1.78, p-value = 0.0127), and others highlighting differences between the same gender from different states (female VIC versus female QLD; pseudo-F = 1.77, p-value = 0.0139). See footnotes on Appendix C: Table C.1 for a complete list of the key differences. When comparing perspective differences in the gender - teaching experience cross, three main differences emerged. Perspectives differed between males and females when comparing individuals with similar years of teaching experience (Appendix C: Table C.1 footnotes: male 1-5years experience versus female 1-5years experience - pseudo-F = 1.74, p-value = 0.0217; male 6-15years experience versus female 6-15years experience - pseudo-F = 1.65, p-value = 0.0278). Male perspectives were found to be different based on years of teaching experience (male 1-5 years' experience versus male 6-15 years' experience - pseudo-F = 1.75, p-value = 0.0122).

Table 3.2 Summary of educator point allocations in Question 1, grouped by demographic factors discipline and education setting. Both the frequency of selection and measures of central tendencies for point allocations are displayed for each statement. Item codes displayed for ease of comparisons to figures.

Statement	Descriptives		Discipline		Education setting			
			Science	History	High school	Tertiary	Science: High school	Science: Tertiary
Critical thinking requires both thinking skills and dispositions, which are mapped against standards.(C_1a)	% chose		7.27	5.56	5.56	7.27	7.25	7.32
	# points	mean	1.75	1.50	1.71	1.67	1.80	1.67
		mode	1	N/A	1	1	1	1
		median	1	1.5	1	1	1	1
		min-max	1 - 4	1 - 2	1 - 4	1 - 3	1 - 4	1 - 3
Critical thinking requires a combination of dispositions, knowledge, and skills, all of which develop over time.(C_1b)	% chose		70.00	69.44	69.44	70.00	75.36	60.98
	# points	mean	3.60	3.68	3.57	3.72	3.58	3.60
		mode	3	5	5	3	3	3
		median	3	4	3	3	3	3
		min-max	1 - 15	1 - 7	1 - 15	1 - 15	1 - 15	1 - 15
Critical thinking consists of higher order thinking skills such as analysis, synthesis and evaluation.(C_1c)	% chose		87.27	86.11	86.11	87.27	86.96	87.80
	# points	mean	2.86	2.36	4.36	4.57	4.25	4.78
		mode	5	5	5	5	5	5
		median	5	5	4	5	4	5
		min-max	1 - 15	1 - 8	1 - 15	1 - 15	1 - 15	1 - 15
Critical thinking is embodied by step-by-step procedures and skill development leading to higher levels of thinking.(C_1d)	% chose		38.18	30.56	30.56	38.18	37.68	39.02
	# points	mean	4.46	4.35	2.69	2.88	2.85	2.88
		mode	3	3	3	3	3	1
		median	3	3	3	3	3	3
		min-max	1 - 9	1 - 5	1 - 7	1 - 9	1 - 7	1 - 9
Standards and criteria are essential elements to critical thinking.(C_1e)	% chose		6.36	13.89	13.89	6.36	5.80	7.32
	# points	mean	3.20	3.31	3.04	3.65	1.25	2.67
		mode	1	1	1	N/A	1	N/A
		median	1	1	1	2	1	2
		min-max	1 - 5	1 - 3	1 - 3	1 - 5	1 - 2	1 - 5
To be effective critical thinkers, students require specific thinking skills and knowledge related to their discipline of study.(C_1f)	% chose		44.55	44.44	44.44	44.55	49.28	36.59
	# points	mean	1.86	1.60	1.44	2.67	3.00	3.67
		mode	3	5	1	5	1	3
		median	3	2.5	3	4	3	3
		min-max	1 - 8	1 - 8	1 - 8	1 - 8	1 - 8	1 - 8
Critical thinking skills should be developed through approaches which are embedded across the curriculum.(C_1g)	% chose		77.27	80.56	80.56	77.27	79.71	73.17
	# points	mean	3.67	3.86	3.64	3.89	3.64	3.70
		mode	5	5	2	5	2	5
		median	3	4	3	4	3	4
		min-max	1 - 15	1 - 8	1 - 15	1 - 8	1 - 15	1 - 7
Critical thinking skills should be developed explicitly and separately from other content matter.(C_1h)	% chose		9.09	2.78	2.78	9.09	7.25	12.20
	# points	mean	1.90	4.00	2.00	2.20	1.60	2.20
		mode	1	N/A	1	2	1	2
		median	2	4	2	2	1	2
		min-max	1 - 3	4 - 4	1 - 4	1 - 3	1 - 3	1 - 3
Critical thinking skills require scaffolding to be transferable in to new contexts.(C_1i)	% chose		30.00	30.56	30.56	30.00	31.88	26.83
	# points	mean	2.79	2.55	2.35	3.62	2.41	3.55
		mode	2	2	2	5	2	5
		median	2	2	2	3	2	3
		min-max	1 - 5	1 - 5	1 - 5	2 - 5	1 - 5	2 - 5
General critical thinking skills are readily transferable to discipline-specific contexts without additional scaffolding.(C_1j)	% chose		33.64	33.33	33.33	33.64	33.33	34.15
	# points	mean	3.08	3.25	3.18	3.00	3.13	3.00
		mode	2	2	2	3	2	3
		median	3	3	3	3	3	3
		min-max	1 - 6	2 - 5	1 - 6	1 - 6	1 - 6	1 - 6
Critical thinking is primarily concerned with the application of the formal rules of logic.(C_1k)	% chose		15.45	13.89	13.89	15.45	14.49	17.07
	# points	mean	2.41	2.60	2.40	2.57	2.30	2.57
		mode	2	1	1	2	3	2
		median	2	1	2	2	2.5	2
		min-max	1 - 6	1 - 5	1 - 5	1 - 6	1 - 3	1 - 6

Table 3.3 Summary of findings for *a priori* hypothesis tests undertaken within the conceptual profile (Question 1 concerning perspectives on the nature of CT), using PERMANOVA to explore the effects of key demographic factors on educator perceptions about CT.

	conceptual sub- profile		
	Q1		
	conceptual perspectives		
factor	pseudo-F	p-value	df
discipline	0.37	0.8442	1
education setting	0.08	0.9689	1
gender	2.97	0.0121	1
state	1.39	0.0915	6
teaching experience	0.18	0.9888	2
discipline vs education setting	0.74	0.5963	1
discipline vs gender	1.18	0.3332	1
discipline vs state*	0.97	0.4980	4
discipline vs teaching experience	0.48	0.8862	2
education setting vs gender	0.88	0.4940	1
education setting vs state*	0.78	0.7217	4
education setting vs teaching experience	0.62	0.7841	2
gender vs state*	1.54	0.0431	6
gender vs teaching experience	2.37	0.0122	2
state* vs teaching experience	1.32	0.1062	8
science - education setting	0.37	0.8530	1
science - gender	2.88	0.0148	1
science - state	1.39	0.0850	6
science - teaching experience	0.11	0.9964	2

*pairwise tests incorporating state as a factor excluded NT and ACT respondents from analysis due to low sample size.

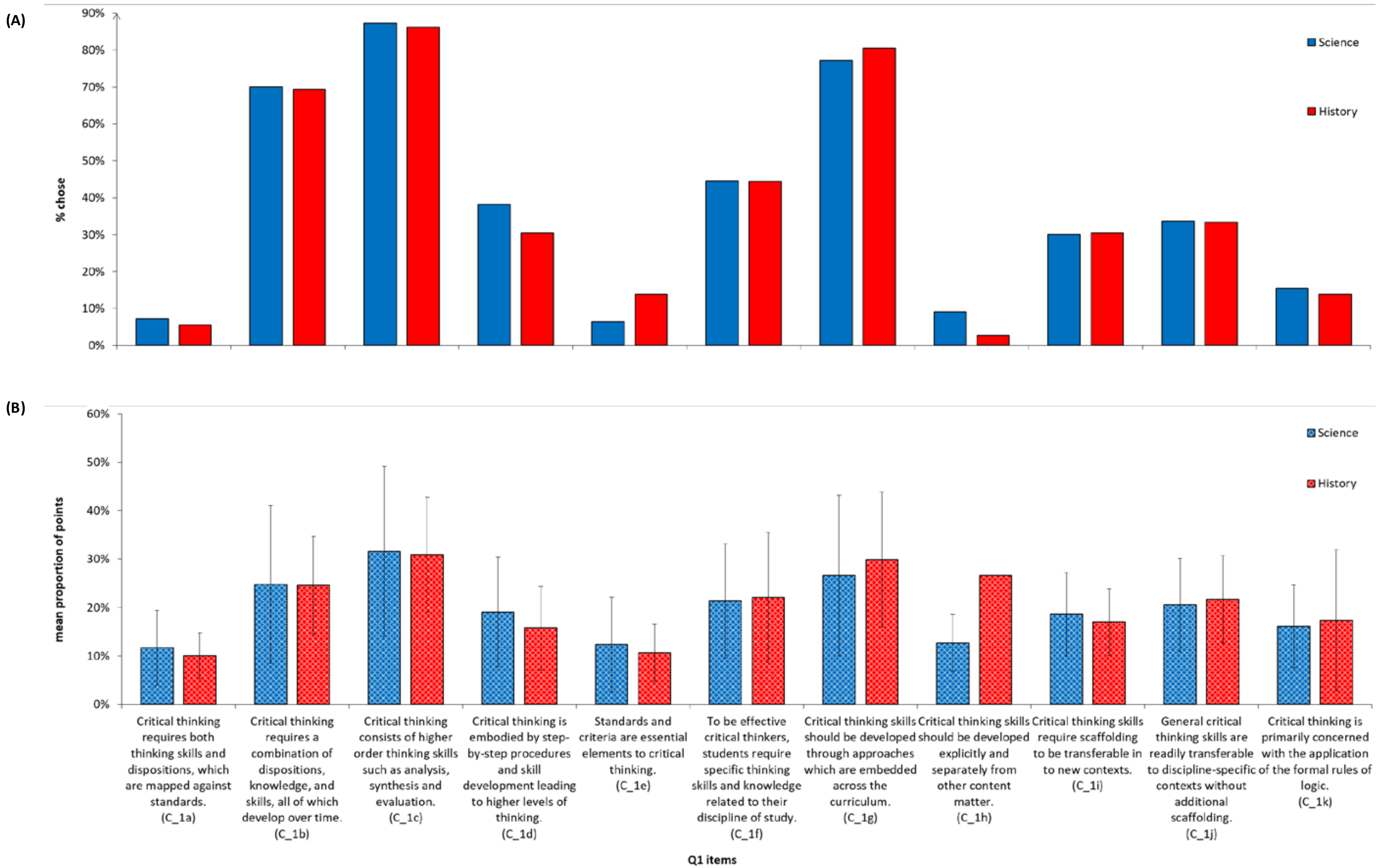


Figure 3.2 A-B Distribution of science and history educator choices (A – solid bars) and point allocations (B – criss-cross bars) for Question 1, comparing the perspectives on the nature of CT. Solid bars represent the % respondents that assigned at least one point to the statement. The criss-cross patterned bar represents the mean proportion (%) of points allocated to the statement by each educator who chose that statement from that group, with the higher the number of points allocated indicating the stronger the statement aligns with their viewpoint. Error bars represent ± 1 SD.

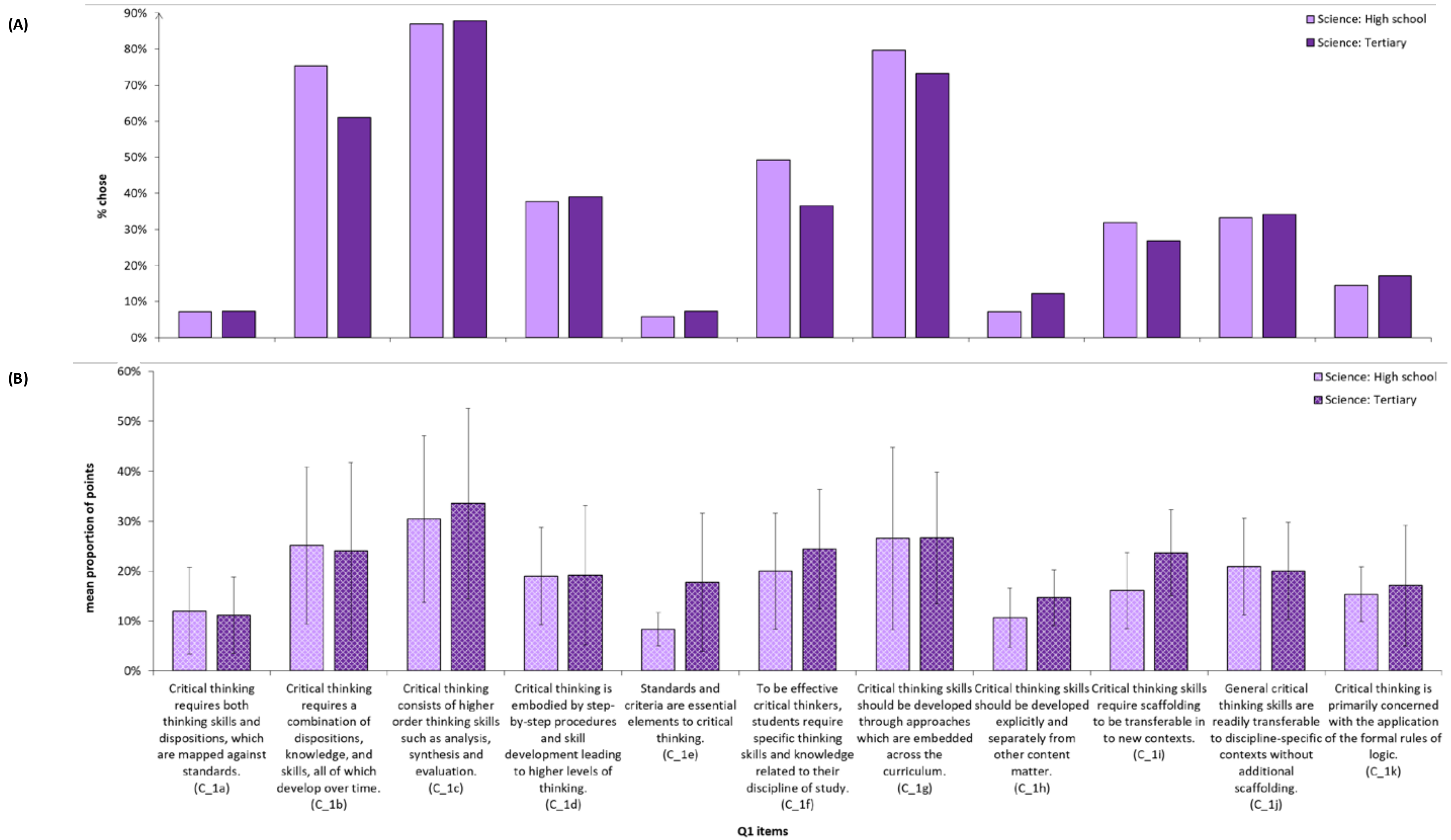


Figure 3.3 A-B Distribution of high school science (light purple) and tertiary science (dark purple) educator choices (A – solid bars) and point allocations (B – criss-cross bars) for Question 1 , comparing the perspectives on the nature of CT. Solid bars represent the % respondents that assigned at least one point to the statement. The criss-cross patterned bar represents the mean proportion (%) of points allocated to the statement by each educator who chose that statement from that group, with the higher the number of points allocated indicating the stronger the statement aligns with their viewpoint. Error bars represent ± 1 SD.

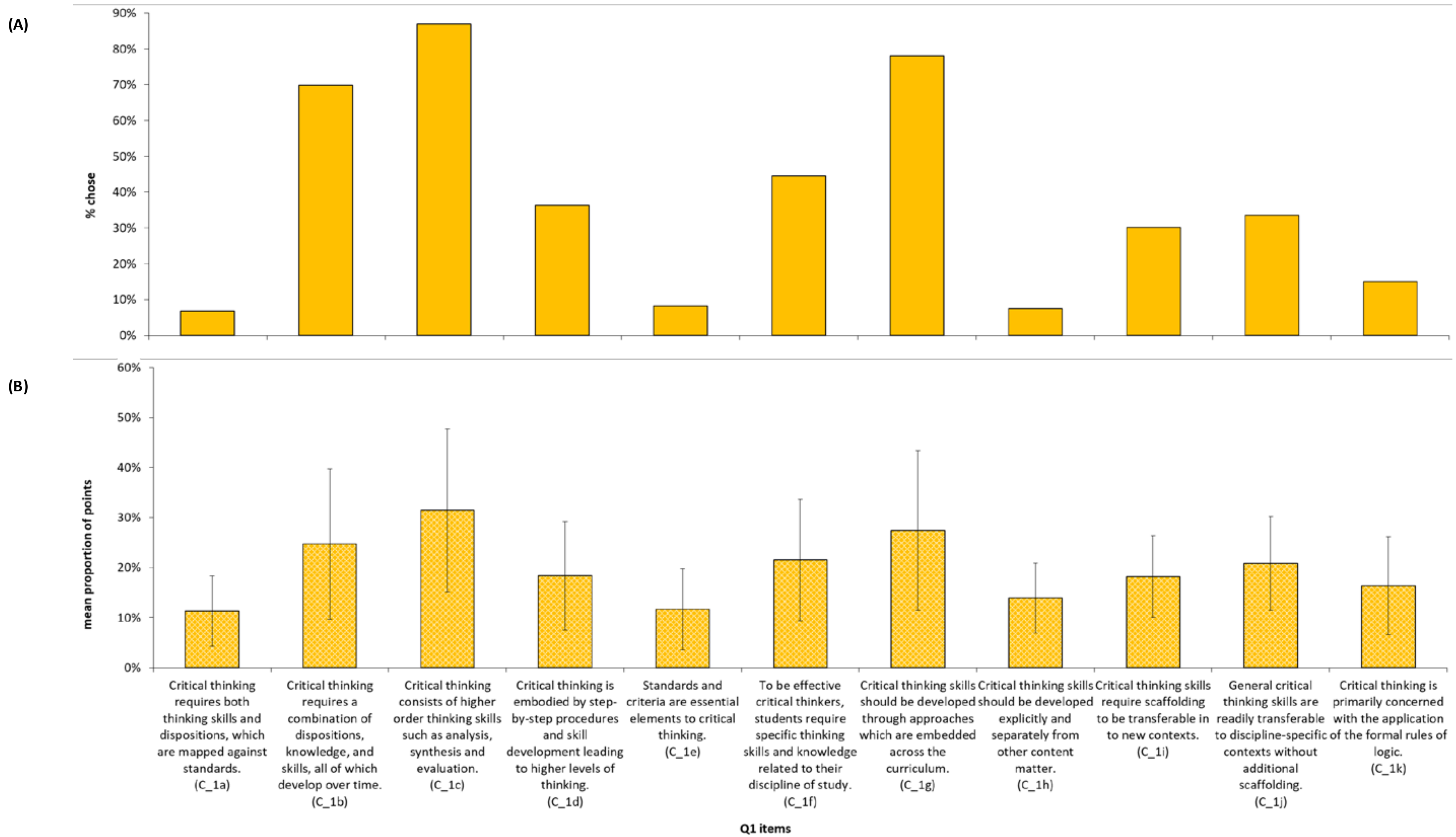


Figure 3.4 A-B Distribution of educator choices (A – solid bars) and point allocations (B – criss-cross bars) to Question 1, comparing the perspectives on the nature of CT. Solid bars represent the % respondents that assigned at least one point to the statement. The criss-cross patterned bar represents the mean proportion (percentage) of points allocated to the statement by each educator who chose that statement from that group, with the higher the number of points allocated indicating the stronger the statement aligns with their viewpoint. Error bars represent ± 1 SD.

Results for Question 1a

Question 1a prompted educators to list features, skills, traits and dispositions they associated with CT. The core themes and terminology used by educators is summarised in a word cloud (Figure 3.5), where analysis, evaluation, logic and questioning stood out as the most frequently included ideas.

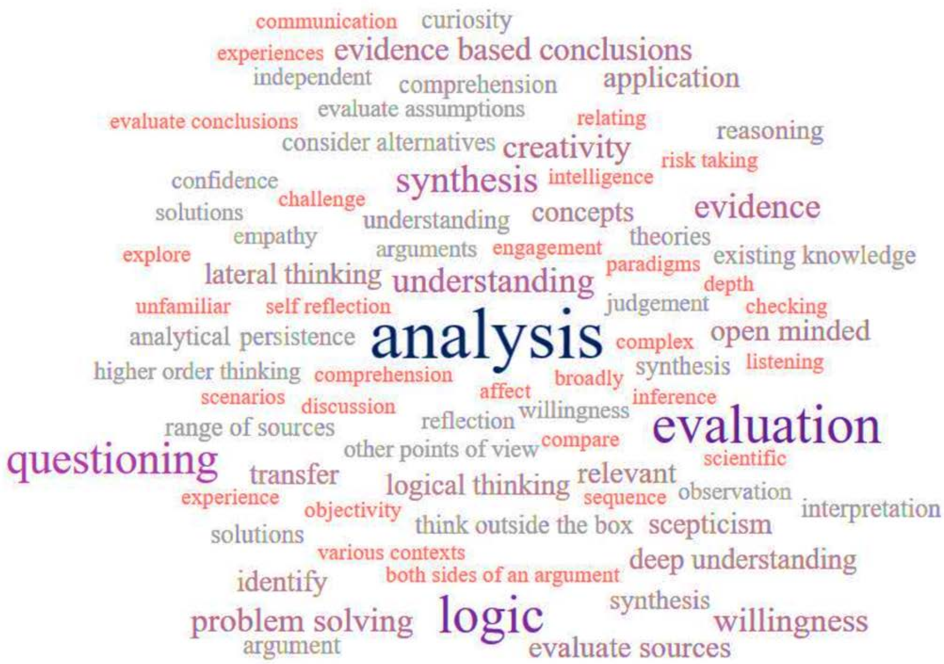


Figure 3.5 Core themes and terms included in educator responses to Question 1a: What features, skills, traits and dispositions do you associate with CT?

Educator responses were also analysed through thematic analysis, where their match to relevant ACT Framework elements and enhancing components was used to code ideas. When examining the entire set of responses, information processing skills were the most commonly referred to aspect of the framework, with 89% of respondents including something related to this element in their response. Applying criteria was the next most commonly included component (47%). However, dispositions were mentioned just as frequently. The inclusions of themes relating to criteria were interesting because educators opted not to spend points on items mentioning standards and criteria in the quantitative question. Purposeful querying was the next most frequently included element in responses, with 45% of educators mentioning themes related to this element in their response. 39% of educators referred to the role of existing knowledge in their response. Themes related to the critical reflection element were included in 33% of respondents; however, only 8% specifically mentioned self-reflection, self-monitoring or metacognition. Lastly, themes around the notion of CT transforming information into understanding were included in 23% of responses.

Table 3.4 summarises the findings concerning the inclusion of ACT Framework themes in educator responses grouped by demographic factors. When comparing the science and history educators, the main difference was the inclusion of some aspect of the critical reflection element in their response, where 34% more history educators included an aspect of this element in their response. Themes around critical reflection were also a point of difference when comparing high school and tertiary educators, as was the inclusion of the element 'applying criteria'. In both instances, tertiary educators were much more likely than high school educators to describe themes relating to these elements in their response (Table 3.4). These trends were upheld in the science-only cohort. However, the margin of difference varied a little more (1-5% comparatively). Aside from themes relating to purposeful querying, where the male respondents (54%) included this theme more frequently than the female respondents (34%), there was generally little difference (<10%) in the inclusion of the framework themes by gender groups.

Considering the sample size differences, there seemed to be little difference in the themes in educator responses by the demographic factor state.[5] However, the South Australian educator responses had a stronger emphasis on themes around CT featuring a transformation of information to understanding (44% included this theme) and purposeful querying (56% included this theme) than responses from educators from other states. Another point of difference was that only 14% Western Australian educators included a response that related to dispositions. This is despite the fact that the question prompted educators to list dispositions they associated with CT as part of their response. So the Western Australian educators stand out from other respondents because of a lack of inclusion of this theme in their extended response.

Concerning teaching experience, those educators with less experience were less likely to include themes relating to purposeful querying, critical reflection and using existing knowledge. However, these educators did include themes relating to CT involving the shift of information to understanding more frequently (39%) and did use terminology relating more to metacognition

[5]Even though educators from the Northern Territory and Australian Capital Territory were included in this analysis; due to sample size (n>5) these trends were not included in these summaries since they may not be representative of the wider perspective of educators from their location.

than the more experienced educators. The main aspect that separated the more experienced educators was that educators who had been teaching the longest (16+ years) were less likely to include themes relating to metacognition and critical reflection compared to the responses from the mid-career educators (5-16 years).

Table 3.4 Results for thematic analysis of Question 1a responses using framework elements for *a priori* groupings.

			framework element					enhancing component		
	Factor	n*	purposeful querying	critical reflection	using information processing skills	applying criteria	using existing knowledge	dispositions	metacognition	information to understanding
discipline	History	32	53%	59%	88%	50%	44%	50%	6%	28%
	Science	100	43%	25%	90%	46%	37%	46%	8%	22%
education setting	High school	89	42%	26%	85%	30%	38%	49%	8%	22%
	Tertiary	43	53%	49%	98%	81%	40%	42%	7%	26%
	Science: High school	66	38%	18%	86%	30%	36%	50%	9%	20%
	Science: Tertiary	34	53%	38%	97%	76%	38%	38%	6%	26%
gender	female	73	38%	33%	90%	44%	38%	42%	10%	26%
	male	59	54%	34%	88%	51%	39%	53%	5%	20%
state	ACT	4	50%	0%	100%	25%	25%	25%	0%	25%
	NSW	41	41%	37%	85%	39%	32%	59%	5%	17%
	NT	1	0%	0%	100%	100%	0%	0%	0%	0%
	QLD	27	44%	19%	89%	44%	33%	52%	7%	30%
	SA	16	56%	50%	88%	69%	50%	56%	6%	44%
	VIC	30	47%	33%	90%	43%	50%	40%	10%	17%
	WA	13	46%	46%	100%	62%	38%	15%	15%	23%
teaching experience	1 to 5 years	23	30%	22%	87%	52%	22%	43%	17%	39%
	6 to 15 years	48	48%	50%	90%	52%	42%	44%	10%	23%
	16+ years	61	49%	25%	90%	41%	43%	51%	2%	18%

*14 of the 146 participants from the study did not answer this question. The sample size for each group has been updated accordingly.

3.4.2 Concerning understanding about CT development

Educators responded to four survey questions (3 quantitative and 1 qualitative) relating to their understanding of the development of CT. The quantitative questions were explored using the PERMANOVA and SIMPER routines at a question level, and as a theoretical set, where the items from three quantitative development questions were incorporated into the development sub-profile. The qualitative question findings are presented as word clouds. The results, contributions of specific question items to group differences and trends associated with each question and the development sub-profile (including by any demographic factors), are explained in the pages that follow.

Findings related to the hypothesis testing undertaken on the quantitative questions and development sub-profile using the PERMANOVA routine.

There were some differences in educator perceptions of development, and a summary of the results from the PERMAOVA routine is presented in Table 3.5. In short, at a question level there were highly significant differences for Question 2 (attributes and outcomes) by education setting (pseudo-F = 4.98, p-value = 0.0002), and for Question 4 (classroom perspectives) by the demographic factors: discipline (pseudo-F = 8.94, p-value = 0.0001), and education setting (pseudo-F = 4.12, p-value = 0.0003). There were also highly significant differences within the development sub-profile when explored by the demographic factors discipline (pseudo-F = 7.14, p-value = 0.0001) and education setting (pseudo-F = 6.02, p-value = 0.0002). However, there were no significant differences identified at the 0.01 level for the demographic factors: state, gender and teaching experience, for any of the quantitative questions individually, nor for the development sub-profile (See Table 3.5). There were also no significant differences by any of the demographic factors in the responses to Question 3 (development approach), nor for any interaction effects at either the question or sub-profile level. Trends from within science cohort were also found to mimic this pattern of significant and non-significant differences (see Table 3.5). Due to the methodology of the PERMANOVA routine, the identity of the individual items driving the key differences between educator groups within the profiles are not readily apparent. However, this information can be garnered at a question level through bar graphs in combination with the SIMPER routine. These trends, as well as the qualitative findings related to Question 3a-c, are explained in the next section.

Table 3.5 Summary of findings for *a priori* hypothesis tests undertaken within the development questions and development sub-profile, using PERMANOVA to explore the effects of key demographic factors on educator perceptions about CT.

	PERMANOVA results for the development sub-profile questions									PERMANOVA results for the developmental sub-profile		
	Q2 attributes and outcomes			Q3 development approach			Q4 classroom perspectives					
factor	pseudo-F	p-value	df	pseudo-F	p-value	df	pseudo-F	p-value	df	pseudo-F	p-value	df
discipline	2.25	0.0479	1	1.27	0.2747	1	**8.94**	**0.0001**	**1**	**7.14**	**0.0001**	**1**
education setting	**4.98**	**0.0002**	**1**	0.21	0.7694	1	**4.12**	**0.0003**	**1**	**6.02**	**0.0002**	**1**
gender	1.20	0.3093	1	0.21	0.7880	1	2.04	0.0369	1	1.15	0.3109	1
state	0.70	0.8593	6	0.64	0.7655	6	1.01	0.4266	6	0.89	0.6432	6
teaching experience	1.32	0.2068	2	0.39	0.7675	2	1.51	0.0929	2	0.97	0.4690	2
discipline vs education setting	0.00	>0.05	1	0.37	0.6297	1	0.84	0.5616	1	0.46	0.8653	1
discipline vs gender	0.89	0.5062	1	1.14	0.2865	1	0.80	0.6116	1	0.76	0.6177	1
discipline vs state*	0.61	0.8643	4	0.63	0.7184	4	0.83	0.6226	4	0.72	0.8414	4
discipline vs teaching experience	0.56	0.8251	2	0.25	0.8747	2	0.96	0.4534	2	0.85	0.5998	2
education setting vs gender	0.89	0.4972	1	0.08	0.8971	1	1.26	0.2672	1	0.91	0.4981	1
education setting vs state*	1.56	0.0645	4	0.89	0.5090	4	1.40	0.0915	4	1.42	0.0810	4
education setting vs teaching experience	0.53	0.8676	2	0.43	0.7348	2	1.46	0.1194	2	0.58	0.8726	2
gender vs state*	1.25	0.2156	6	1.00	0.4171	6	1.34	0.1211	6	1.37	0.0756	6
gender vs teaching experience	1.12	0.3454	2	0.84	0.4635	2	0.82	0.6550	2	0.79	0.6700	2
state* vs teaching experience	0.81	0.7576	8	0.86	0.5847	8	1.18	0.2040	8	0.72	0.9234	8
science - education setting	**4.16**	**0.0013**	**1**	0.21	0.7641	1	**4.10**	**0.0002**	**1**	**5.36**	**0.0003**	**1**
science - gender	0.54	0.7608	1	0.87	0.3676	1	2.18	0.0232	1	1.22	0.2749	1
science - state	0.69	0.8687	6	0.79	0.6426	6	0.95	0.5163	6	0.81	0.7857	6
science - teaching experience	0.92	0.5221	2	0.17	0.9589	2	1.47	0.1039	2	0.87	0.5787	2

*pairwise tests incorporating state as a factor excluded NT and ACT respondents from analysis due to low sample size.

Trends for Question 2

Question 2 asked educators to compare CT's value as a learning outcome against other learning attributes and outcomes. At a question level, perspectives relating to the value of CT as a learning outcome only differed significantly by education setting (refer back to Table 3.5). According to the SIMPER analysis, responses to items D_2b (citizenship) and D_2g (intercultural understanding) provided the greatest difference between high school and tertiary educators perspectives (Figure 3.6). For citizenship, high school educators were split across the categories: equal importance (41%) and CT more important (50%) options; whereas for the tertiary educators, the CT more important option received 76% of the group allocation (Table 3.6). There was a similar pattern for intercultural understanding. Again, more tertiary educators indicated that they valued CT more as a learning outcome (56%), whereas high school were split between equal importance (47%) and CT having more importance (44%) than intercultural understanding (see Table 3.6). Item D_2b (citizenship) was also the highest discriminating item contributing to differences between science respondents by their education setting (Figure 3.6). However, when factoring in the mean scores and standard deviations for the second item (as per the SIMPER test approach), item D_2e (ethical conduct) contributed more to group differences than D_2g (intercultural understanding) (Figure 3.12). In this instance, there was a three-way split on perspectives within both education settings; however, the tertiary educators were slightly more likely to indicate that CT has equal to more importance than ethical conduct (78%) than the high school educators (70%).

Even though there were no significant differences according to the PERMANOVA approach, when exploring this question by the demographic factor discipline, there were also some noteworthy trends for particular items. Perspectives by discipline varied the most on views towards items D_2g and D_2j. For example, 56% of science educators indicated intercultural understanding (item D_2g) had lower importance than CT, whereas history educators gave this as equally important, with only 22% of history educators rating CT more important (Table 3.6). For item D_2j (numeracy), history educators indicated that numeracy had lower importance than CT (47%) while science educators generally indicated CT having equal importance, with only 14% of science educators ranking CT as more important that numeracy (Table 3.6). Teamwork was another item where perspectives differed by educator discipline. The majority of history educators viewed CT as more important than CT (64%), whereas the science educators were more evenly split between the 'CT has more importance' and 'CT has equal importance' categories (44% versus 42%, respectively). These differences were likely masked in the multivariate question analysis (comprising of 15 items) since both science and history educator groups predominantly rated 9 out of the 15 items as having equal importance between the learning outcome and CT (see Appendix C: Table C.3).

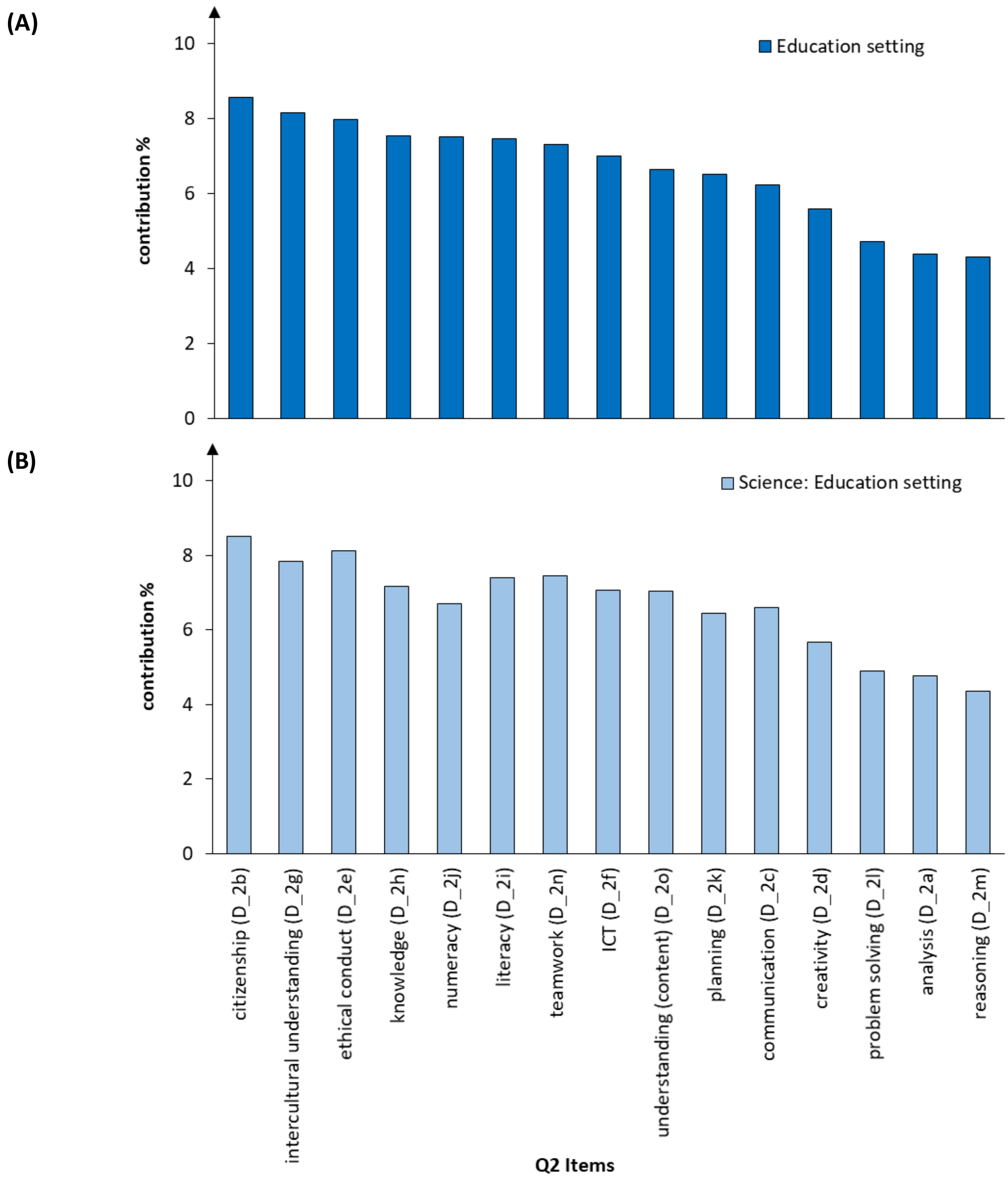


Figure 3.6 A-B Percentage contribution for each Question 2 survey items to the differences between demographic factor groupings from education setting (A) and science: education setting (B). Displayed from highest to lowest contribution based on SIMPER analysis of education setting differences. Demographic factor and education setting were identified as significant through hypothesis testing.

Table 3.6 Proportions of educators who indicated agreement with corresponding Likert-type scale option for Question 2 items 2b, 2e, 2g, 2j and 2n for key demographic factors: discipline and education setting (overall and within the science cohort). Proportions for all Question 2 items are found in Appendix C: Table C.3

			Proportions by demographic factors#					
			Discipline		Education setting		Science: Education setting	
Item	Item description	Options	Science (n=110)	History (n=36)	High school (n=96)	Tertiary (n=50)	High school (n=69)	Tertiary (n=41)
D_2b	citizenship	CT less important (1 -2)	7%	6%	9%	2%	10%	2%
		equal importance (3)	35%	33%	41%	22%	42%	22%
		CT more important (4-5)	58%	61%	50%	76%	48%	76%
D_2e	ethical conduct	CT less important (1 -2)	27%	22%	29%	20%	30%	22%
		equal importance (3)	46%	58%	48%	52%	45%	49%
		CT more important (4-5)	26%	19%	23%	28%	25%	29%
D_2g	intercultural understanding	CT less important (1 -2)	8%	17%	9%	12%	6%	12%
		equal importance (3)	35%	61%	47%	32%	41%	27%
		CT more important (4-5)	56%	22%	44%	56%	54%	61%
D_2j	numeracy	CT less important (1 -2)	32%	14%	31%	20%	36%	24%
		equal importance (3)	55%	39%	51%	50%	55%	54%
		CT more important (4-5)	14%	47%	18%	30%	9%	22%
D_2n	teamwork	CT less important (1 -2)	15%	8%	19%	2%	22%	2%
		equal importance (3)	42%	28%	42%	32%	45%	37%
		CT more important (4-5)	44%	64%	40%	66%	33%	61%

due to rounding to whole numbers proportions may be ± 1% of the sum of the displayed values.

Trends for Question 3

Question 3 asked educators to select their CT development philosophy, with options including an embedded development approach, an explicit development approach, and no planned approach. Overall, the majority of educators indicated they used an embedded approach for CT development (Figure 3.7), which is consistent with recommendations with Australian education policies related to CT. However, there was a subset of around 38% of educators, who indicated they use an explicit approach to develop CT (Figure 3.7).

Examination of the frequencies of educator choices for each category by the demographic factor education setting revealed that the proportions of high school and tertiary educators were approximately equal within any one category (Table 3.7). The majority of educators indicated that their CT development philosophy was based on an embedded approach. This same trend is evident when exploring within the science education cohort. The similarity in educator choices by education setting was confirmed by the PERMANOVA routine, where there were no significant differences by the demographic factor education setting (Table 3.5). Exploration by discipline revealed there were similar proportions of science and history educators who used an embedded approach (item D_3b) for developing CT. However, a higher proportion of history educators indicated they used an explicit development approach (D_3a) (Table 3.7). As stated at the start of Section 3.4.2, there was not a significantly different response pattern between science and history educators (refer back to Table 3.3), nor for any of the other demographic factors.

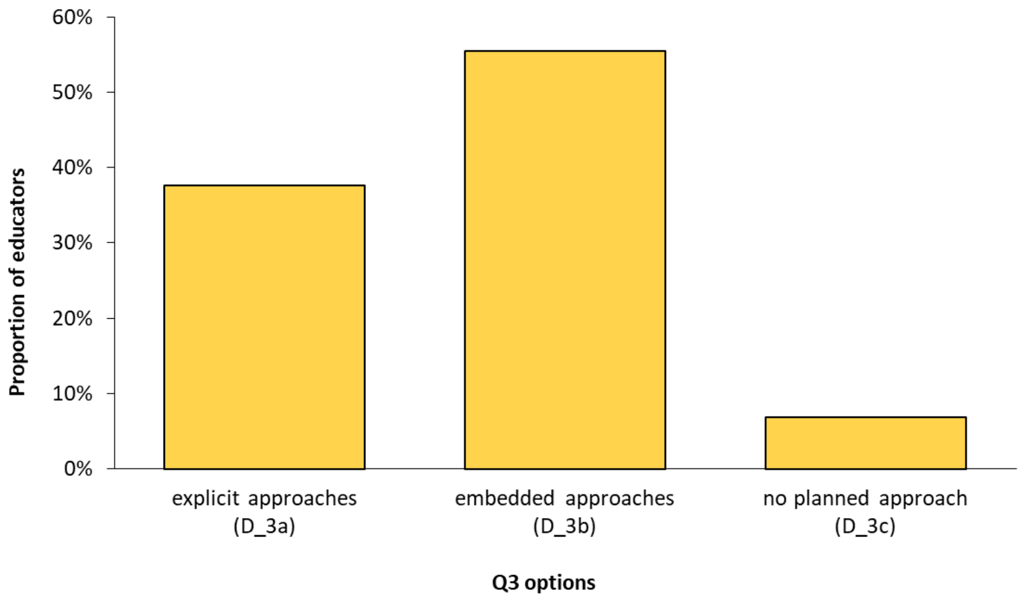


Figure 3.7 Proportions of educators per option for Question 3, concerning the educator's philosophy for CT development.

Table 3.7 Proportions of educators who chose each option in Question 3. This question asked educators about their plan for CT development.

Factor		Options for Q3: Plan involving explicit instruction (D_3a)	Plan involving approaches embedded across the curriculum (D_3b)	No plan (D_3c)
Discipline	Science (n=110)	35%	56%	9%
	History (n=36)	47%	53%	0%
Education setting	High school (n=96)	39%	56%	5%
	Tertiary (n=50)	36%	54%	10%
Science: Education setting	High school (n=69)	36%	57%	7%
	Tertiary (n=41)	32%	56%	12%

Results for Question 3a, 3b and 3c

For these qualitative questions, educators were asked to expand on their response to quantitative Question 3 (explicit, embedded or no plan) by explaining more about their approach to developing CT. 128 of the 146 educators responded. Educator responses included a mix of ideas about their expectations and observations of students, specific activities and tools they incorporate in their classrooms, and some mentioned the types of assessment tasks they use.

Educators who selected the 'no plan' option (n = 8) were asked how they knew if CT skills were an emergent property of your students' learning. All respondents who picked this option were science educators. Of those that provided a qualitative response (n = 7), explanations of their understanding about student CT development generally included knowing from classroom activities and assessments (i.e. setting evaluation tasks such as evaluation of evidence, claims, readings), and knowing from interactions with the students (i.e. can notice it in the question students asked or from conversations and discussions). However, some respondents also mentioned that they "don't formally know" since they "don't directly assess it," or that it is "incidental" in the other tasks they set (Appendix C Table C.4).

For educators that selected explicit development in Question 3 (n = 52), 49 explained at least one specific action to develop CT. Having students answer questions and work on various assessment tasks were the most common themes (Figure 3.8), there was also a focus on using analysis skills to examine experimental results, experimental designs and examining sources for quality and bias. The themes were similar to the responses obtained from the cohort who responded to the embedded question; however, the language used was more specific. These educators were also asked if they assess their approach, and if they had any further comments. 32% of educators (n=16) indicated they had thought about assessing, and 38% educators (n=19) indicated they had tried assessing it. One respondent specifically commented that they were unsure how to assess CT properly. Another noted that even though they have assessed their approach, it is difficult to do so. Only one responded described the use of a ‘pre-test, post-test’ assessment format, which was a part of the development program they use (Think Science Australia program).

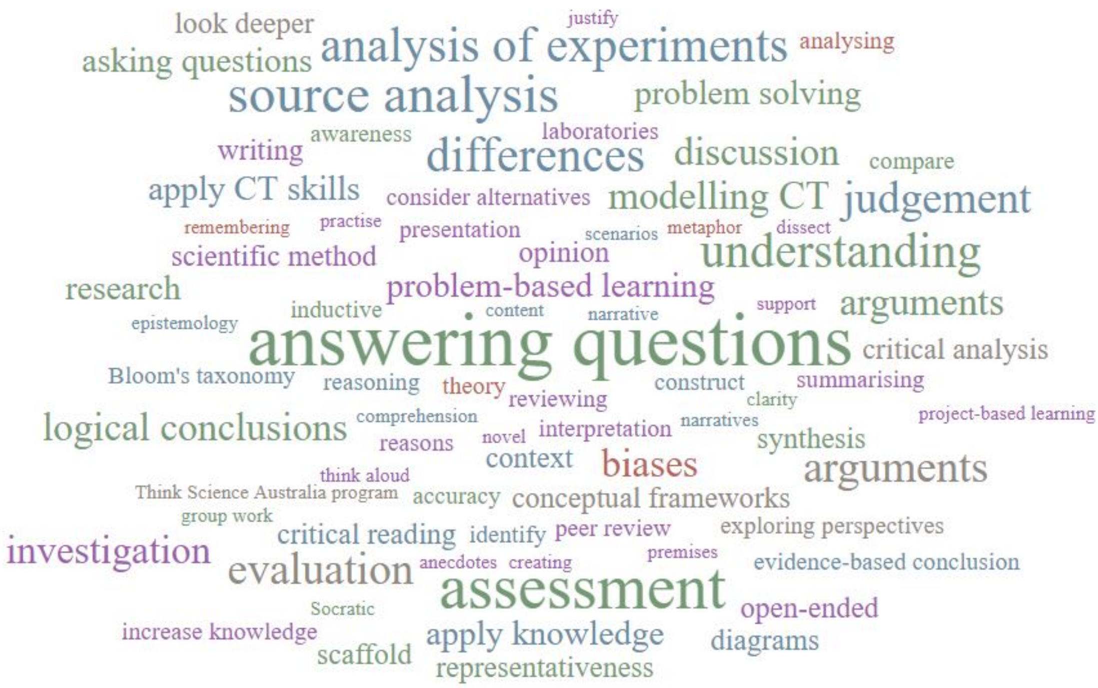


Figure 3.8 Core themes and terms included in educator responses to Question 3a: Please provide an example of the program or approach you use to develop your students' CT skills.

For educators that selected embedded development in Question 3 (n = 72), 71 gave explanations about their perceptions of their approach for developing CT. Most respondents listed an activity or approach they use with students, but only 22 (31%) used terminology indicating some degree of formal evaluation (such as assignments, assessments, reports and tasks). These were grouped under the theme assessment in the word cloud. As indicated through the frequency-based word cloud in Figure 3.9, themes around asking students questions, students asking questions, extending student thinking, using problem-solving and modelling CT were most frequently mentioned. Even though educators had previously indicated they used an embedded approach to develop CT, six of the high school educators (three science, and three history) mentioned providing explicit CT instruction in their response.

Figure 3.9 Core themes and terms educators included in their response to Question 3b: How do you ensure that CT is developed in your students by this embedded approach?

Trends for Question 4

For addressing perspectives on different classroom approaches (Question 4), educators indicated agreement with Likert-type statements concerning the effectiveness of a particular classroom approach for developing CT. The effectiveness of each classroom approach option was informed by at least 50% of respondents (Figure 3.10). Twelve of the eighteen approaches were rated as effective for developing CT by at least 50% of the educators who have employed that method at some point in their teaching career (Figure 3.11). Problem-based learning was rated as the most effective way to develop CT (87% indicated it was effective), and oral presentations received the least amount of support as an effective way to develop CT (only 21% of educators rated it as effective for developing CT). There were differing perspectives about the effectiveness of peer assessment for developing CT (D_4i). This item received the highest proportion of educators in the 'ineffective' category (36%), but equally as many educators also indicated it was 'effective' for developing CT (38%).

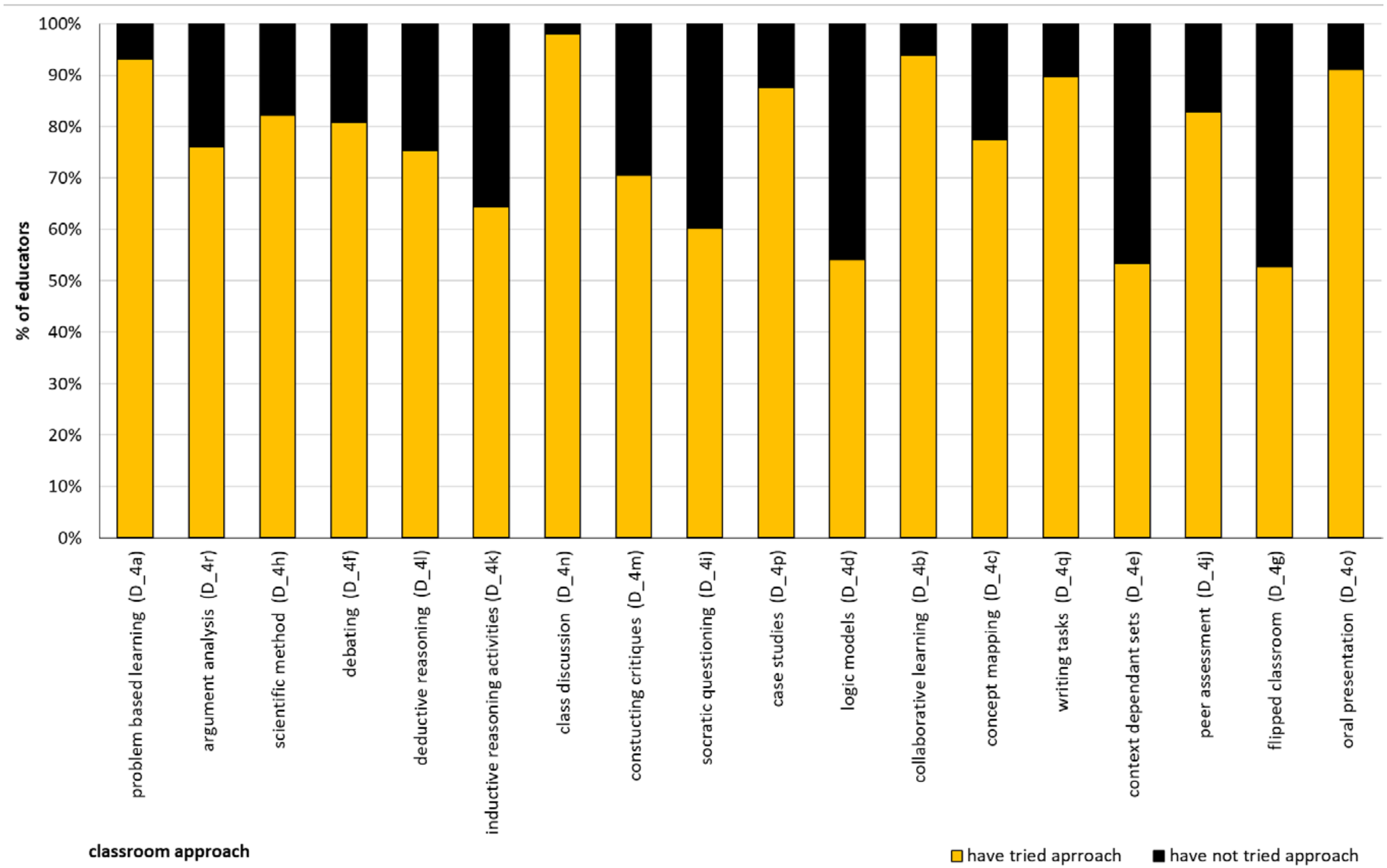


Figure 3.10 Summary of experience with each of the classroom approaches included in Question 4.

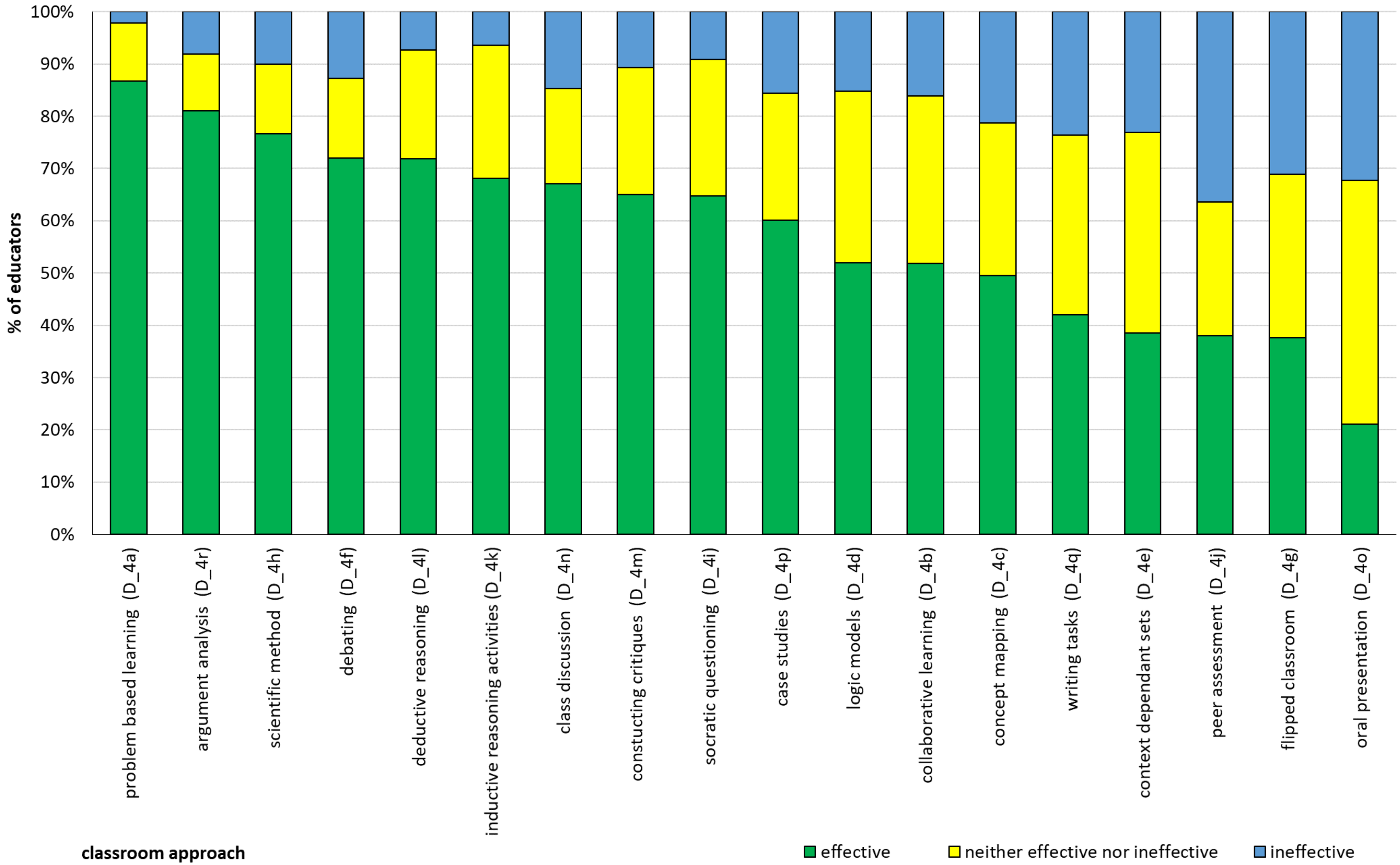


Figure 3.11 Summary of the effectiveness of each classroom approach for developing CT from Question 4 , based on the perspectives of educators who have had experience with this approach.

For Question 4, there were some significant differences in educator perspectives by the demographic factors: discipline (Table 3.5: pseudo-F = 8.94, p-value = 0.0001), and by education setting (Table 3.5: PERMANOVA for education setting: pseudo-F = 4.12, p-value = 0.0003; and PERMAVONVA for science: education setting: pseudo-F = 4.10, p-value = 0.0002). SIMPER was again used to explore which particular items were contributing the most to differences between the educator groups. These analyses highlighted that different items were driving the key differences by discipline and education setting. For the difference by discipline, the best discerning item was D_4h (Figure 3.12, Table 3.8), which addressed the effectiveness of the scientific method for developing CT. Where frequency analysis revealed that science educators generally rated the scientific method item more highly (88%) than history educators (11%) – see Table 3.9. However, the trend reversed for the second highest contributing (item D_4i), with history educators generally rating the effectiveness of Socratic questioning (D_4i) for developing CT more highly (58%) than the science educators (33%).

The best discerning items contributing to differences between high school and tertiary educator perspectives was not as clear, with eight items contributing similar amounts (>6% each) to the overall differences between the high school and tertiary educators (Figure 3.10 A, and Table 3.9). The same eight items contributed the most to group differences for analysis by education setting and by science: education setting; however, the order of ranks 4-8 was different in the science educator cohort (Table 3.9). For example, Socratic questioning (D_4i) had less of an impact on group difference in the science only cohort (rank 8) compared to perspectives when the history educators were present in the response set (i.e. education setting: rank 4). This is not surprising given that this item was identified as the second most distinguishing item between group differences when exploring group differences by discipline, indicating a discipline-based preference among the history educators. While there was some variation in the top eight, the top three contributing items were the same for group differences by education setting, irrespective of the presence of the history educators in the response set (see Figure 3.13 A-B, and Table 3.9 items: D_4c (concept mapping); D_4r (argument analysis); and D_4k (inductive reasoning activities)). High school educators generally rated both concept mapping, inductive reasoning activities and argument analysis as effective for developing CT (43%; 45%; and 65%; respectively); whereas tertiary educators generally had less experience with these development approaches, and between 9-13% fewer tertiary educators thought that they were effective (Table 3.8). However, the difference between the number of high school and tertiary educator perspectives who indicated inductive reasoning was effective for developing CT was more similar (44% versus 42% respectively, see Table 3.8). These trends were maintained in the science-only cohort, with the tertiary educator again having less experience with the concept mapping, inductive reasoning and argument analysis approaches (refer to n/a category Table 3.8), and there generally being a higher proportion of high school science educators than tertiary science educators indicating the approach as effective for developing CT (Table 3.8). A summary of all the proportions for Question 4 can be found in Appendix C: Table C.5.

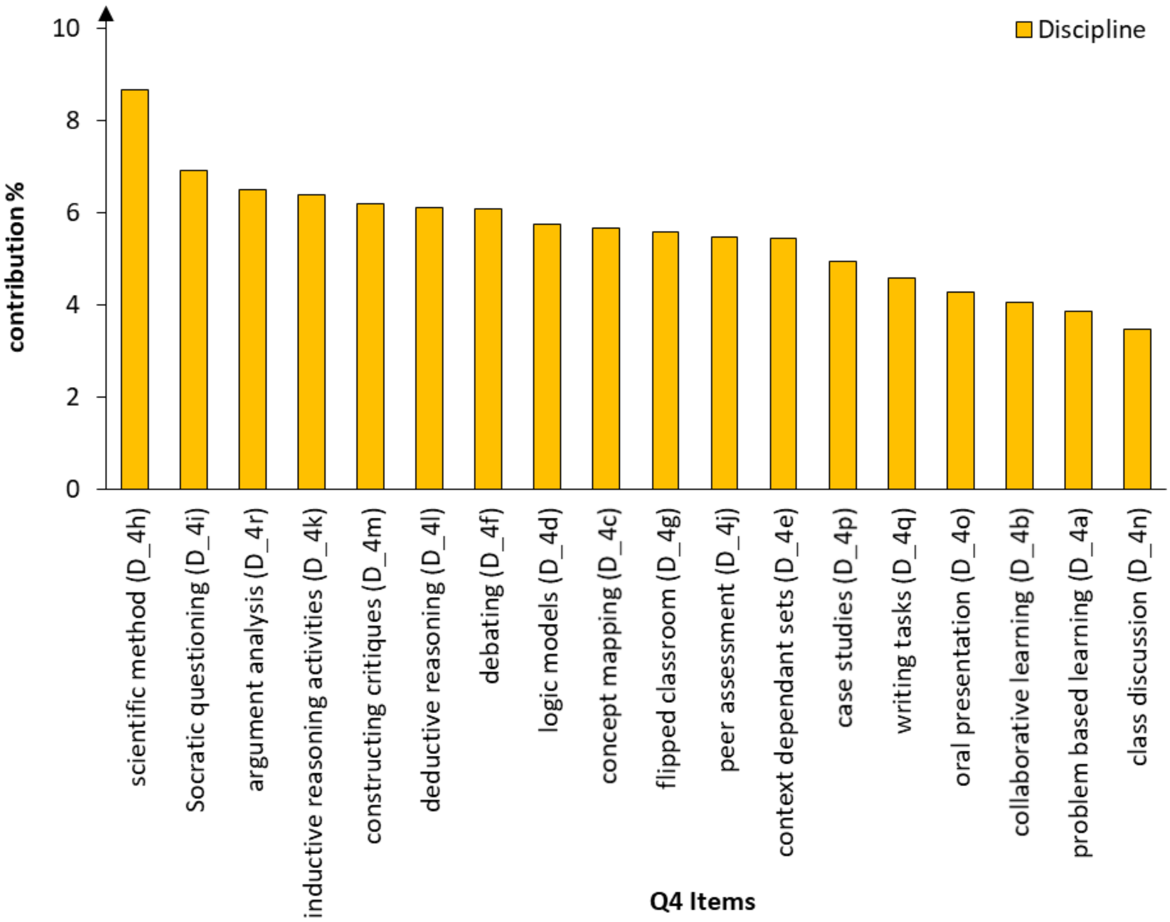


Figure 3.12 Percentage contributions for Question 4 survey items to the differences between educators by discipline. This demographic factor was identified as significant through hypothesis testing. Displayed from highest to lowest contribution based on SIMPER analysis to help reveal which items contributed most to group differences.

Table 3.8 Proportions of educators who indicated agreement with corresponding Likert-type scale option for Question 4 items 4c, 4h, 4i, 4k and 4r, for key demographic factors: discipline, education setting, and also by education setting within the science cohort. Proportions for all Question 4 items can be found in Appendix C: Table C.5.

			Proportions by demographic factors					
			Discipline		Education setting		Science: Education setting	
Item	Item description	Options	Science (n=110)	History (n=36)	High school (n=96)	Tertiary (n=50)	High school (n=69)	Tertiary (n=41)
D_4c	concept mapping effectiveness	ineffective	15%	22%	23%	4%	22%	2%
		neither effective or ineffective	21%	28%	27%	14%	25%	15%
		effective	39%	36%	43%	30%	48%	24%
		no experience	25%	7%	7%	52%	5%	59%
D_4h	scientific method effectiveness	ineffective	5%	17%	6%	12%	4%	7%
		neither effective or ineffective	9%	17%	10%	12%	6%	15%
		effective	80%	11%	68%	54%	88%	66%
		no experience	6%	55%	16%	22%	2%	12%
D_4i	Socratic questioning effectiveness	ineffective	5%	6%	8%	0%	9%	0%
		neither effective or ineffective	17%	11%	17%	14%	17%	17%
		effective	33%	58%	40%	38%	33%	32%
		no experience	45%	25%	35%	48%	41%	51%
D_4k	inductive reasoning activities	ineffective	5%	3%	6%	0%	7%	0%
		neither effective or ineffective	17%	14%	21%	8%	23%	7%
		effective	45%	39%	45%	42%	46%	44%
		no experience	33%	44%	28%	50%	24%	49%
D_4r	argument analysis effectiveness	ineffective	7%	3%	8%	2%	10%	2%
		neither effective or ineffective	8%	8%	10%	4%	10%	5%
		effective	55%	83%	65%	56%	58%	49%
		no experience	30%	94%	17%	38%	22%	44%

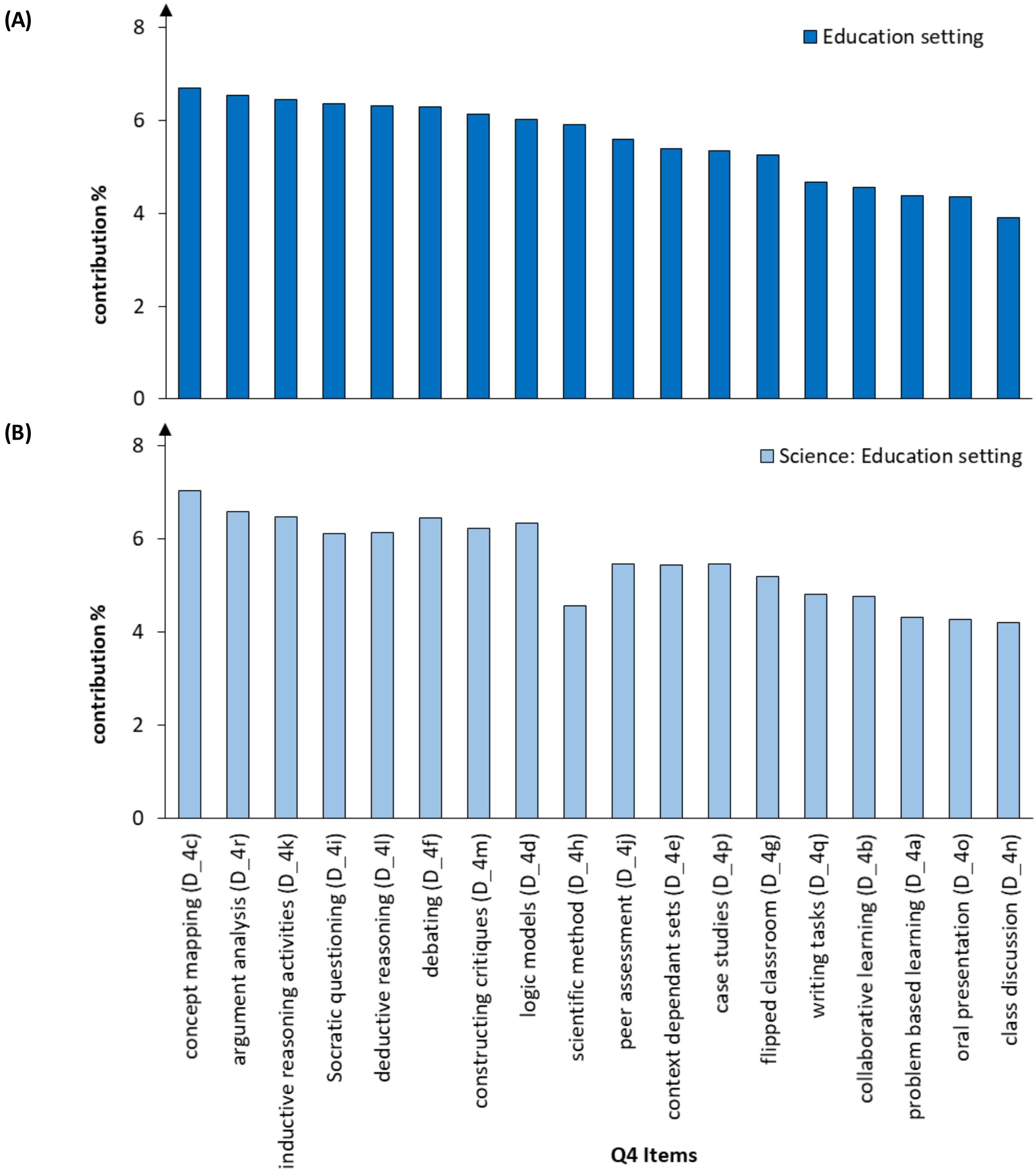


Figure 3.13 A-B Percentage contribution for each Question 4 survey item to the differences between demographic factor groupings from education setting (A) and science: education setting (B). These demographic factor groupings were identified as significant through hypothesis testing. Displayed from highest to lowest contribution based on SIMPER analysis of education setting differences.

Table 3.9 Summary of group dissimilarity analysis from SIMPER approach for Question 4. Displayed in rank order of contributions to the differences between high school and tertiary educators. Other demographic factors (discipline and science: education setting) displayed for comparisons of the items driving group difference across the three demographic factors, which were identified as significantly different using the PERMAONVA approach.

Classroom approach items	item code	SIMPER rank (% contribution of item to difference between groups; group with higher mean score)					
		Discipline (Sc v His)		Education setting (HS vs TER)		Education setting (Sc HS v Sc TER)	
concept mapping	D_4c	9	(5.68%; His)	1	(6.68%; HS)	1	(7.04%; Sc HS)
argument analysis	D_4r	3	(6.51%; His)	2	(6.54%; HS)	2	(6.6%; Sc HS)
inductive reasoning activities	D_4k	4	(6.39%; Sc)	3	(6.43%; HS)	3	(6.49%; Sc HS)
Socratic questioning	D_4i	2	(6.92%; His)	4	(6.34%; HS)	8	(6.11%; Sc HS)
deductive reasoning	D_4l	6	(6.12%; Sc)	5	(6.3%; HS)	7	(6.14%; Sc HS)
debating	D_4f	7	(6.08%; His)	6	(6.28%; HS)	4	(6.46%; Sc HS)
constructing critiques	D_4m	5	(6.19%; His)	7	(6.12%; HS)	6	(6.24%; Sc HS)
logic models	D_4d	8	(5.74%; Sc)	8	(6.01%; HS)	5	(6.35%; Sc HS)
scientific method	D_4h	1	(8.67%; Sc)	9	(5.91%; HS)	15	(4.58%; Sc HS)
peer assessment	D_4j	11	(5.47%; His)	10	(5.58%; HS)	10	(5.47%; Sc HS)
context dependant sets	D_4e	12	(5.44%; Sc)	11	(5.39%; HS)	11	(5.44%; Sc HS)
case studies	D_4p	13	(4.95%; His)	12	(5.33%; HS)	9	(5.47%; Sc HS)
flipped classroom	D_4g	10	(5.58%; His)	13	(5.26%; HS)	12	(5.19%; Sc HS)
writing tasks	D_4q	14	(4.58%; His)	14	(4.67%; TER)	13	(4.81%; Sc TER)
collaborative learning	D_4b	16	(4.07%; His)	15	(4.55%; HS)	14	(4.78%; Sc HS)
problem based learning	D_4a	17	(3.85%; Sc)	16	(4.37%; HS)	16	(4.33%; Sc HS)
oral presentation	D_4o	15	(4.29%; His)	17	(4.35%; HS)	17	(4.28%; Sc HS)
class discussion	D_4n	18	(3.48%; His)	18	(3.9%; HS)	18	(4.22%; Sc HS)

Trends for development sub-profile

The ordination of the development profile (Figure 3.14) demonstrates the respondents separate into two clusters with a few outliers. As stated at the beginning of Section 3.4.2, there were some significant differences in educator perceptions of development evident in the development sub-profile (refer back to Table 3.5). However, exploration of the respondents in each of these clusters revealed a mix of the significant demographic factors contained within each cluster (Figure 3.15 and Figure 3.16). Even though there were significant differences between certain demographic factors, something else was contributing highly to trends within educator groups. Exploration of the contributions of items to group dissimilarity at the sub-profile level using the SIMPER routine revealed that Question 4 contributed the most to group differences for each significant demographic factor (Table 3.10). While some of this is attributable to this question contributing the most items to the sub-profile (18 of the 36 items), exploration of the min-max contributions of items for each question revealed that all 18 items contributed more to the sub-profile than any of the 15 items from Question 2. In fact, perspectives about the scientific method (theme of item D_4h) contributed the most to differences in the science and history educator comparison (Table 3.11). In further support of the fact that is not just the number of items that determine the importance of a question to the sub-profile, Question 3 had the fewest items (n=3), yet also had two items that contributed highly to group differences. For example, items D_3a (explicit approaches) and 3b (embedded approaches) consistently ranked in the top 3 contributors to group differences for each significantly different demographic factor exploration – Table 3.11). Further, when exploring sub-profile differences by education setting, one item from Question 3 (D_3b: embedded approaches) contributed more to group differences than any other individual item in the sub-profile (Table 3.11). Item D_3b also contributed the most to differences within the science educator cohort. However, the rank order of the other items contributing to group differences across the two education setting factors (such as high school versus tertiary; and science: high school versus science: tertiary) was different.

Lastly, since the SIMPER results revealed that 2 out of the 3 items from Question 3 were contributing highly to group differences, it was plausible that this question was the variable separating the two main clusters pictured in the ordination. In fact, labelling the data points by an educator's response to Question 3 shows the strong influence of this question on the shape of the data cloud (Figure 3.17). This ordination clearly demonstrates that there were patterns in developmental perspectives, where educators who chose a particular option in Question 3 (either D_3b or D_3a) also responded similarly on other items in the sub-profile. In contrast, the data points for respondents who chose the no plan option for Question 3 (D_3c) were randomly dispersed throughout the data cloud, indicating fewer similarities between these educators in terms of their responses across items in the development sub-profile, even though they gave the same response to Question 3. The ordination also demonstrates that the set of educators responded differently (across the sub-profile), to educators who chose one of the other options for Question 3. Because if they had answered more similarly, they would be closer together, or even mixed, in the data cloud (as seen with using demographic factors as a way to identify patterns in perspectives in the cloud – see Figure 3.15 and 3.16).

Collectively, these findings at an individual question and sub-profile level indicate variability in the way CT development is approached, as well as patterns in the reported experiences and perspectives of educators who approach CT development in a particular way. Thus, demonstrating the importance of considerations about the nature of CT development (explicit versus embedded), as well as educator demography when forming policies or decisions about CT, as both influence the choices about CT development.

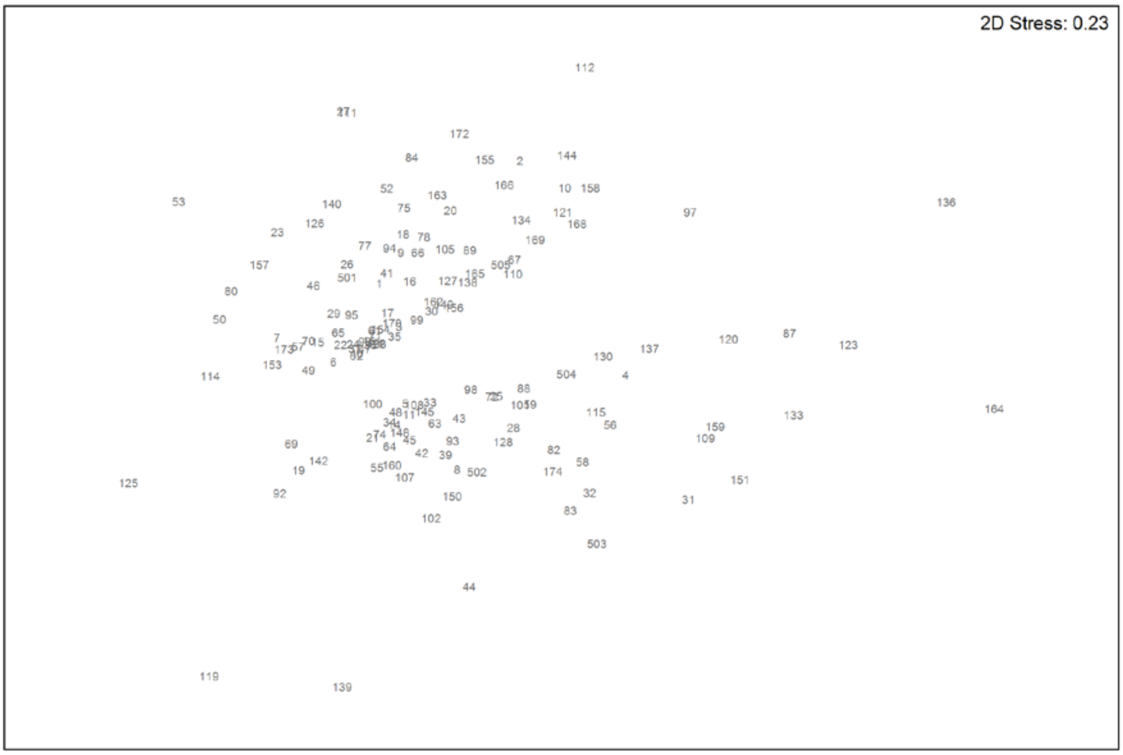


Figure 3.14 Two-dimensional MDS plot of the development sub-profile of educator perspectives on CT (n=146).

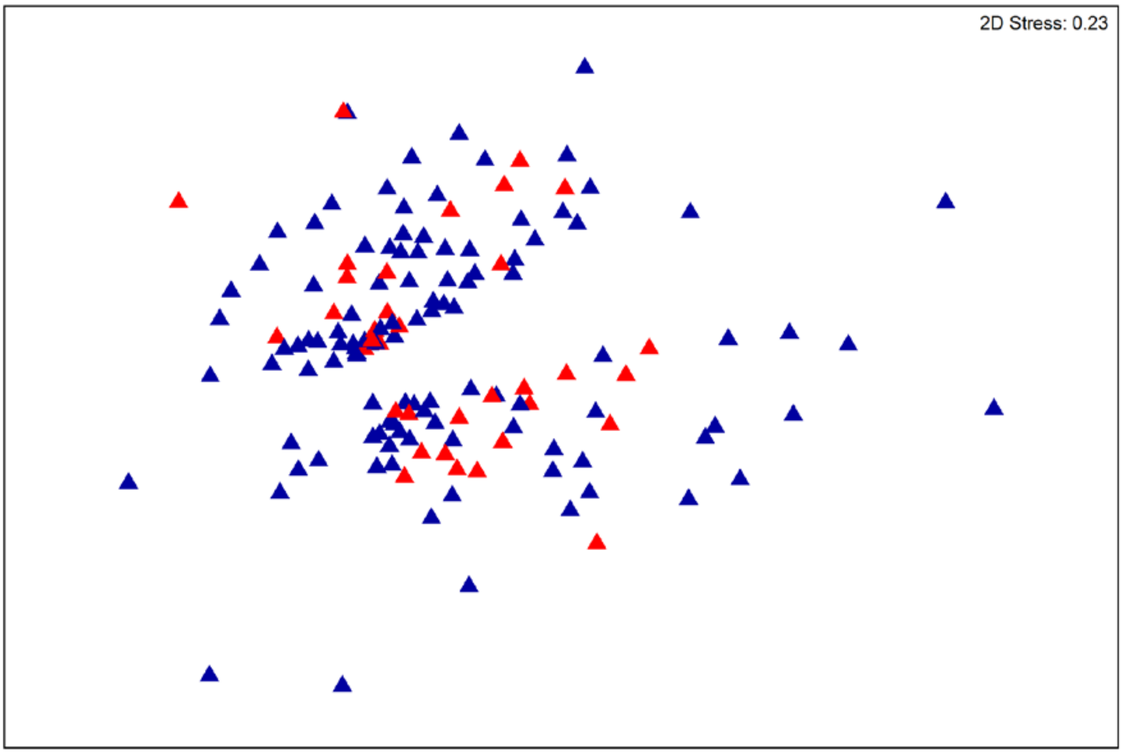


Figure 3.15 Two-dimensional MDS plot of the development sub-profile of educator perspectives on CT (n=146).Educators are coloured by the education setting they teach in, with the blue triangles representing the science educators and the red triangles representing the history educators.

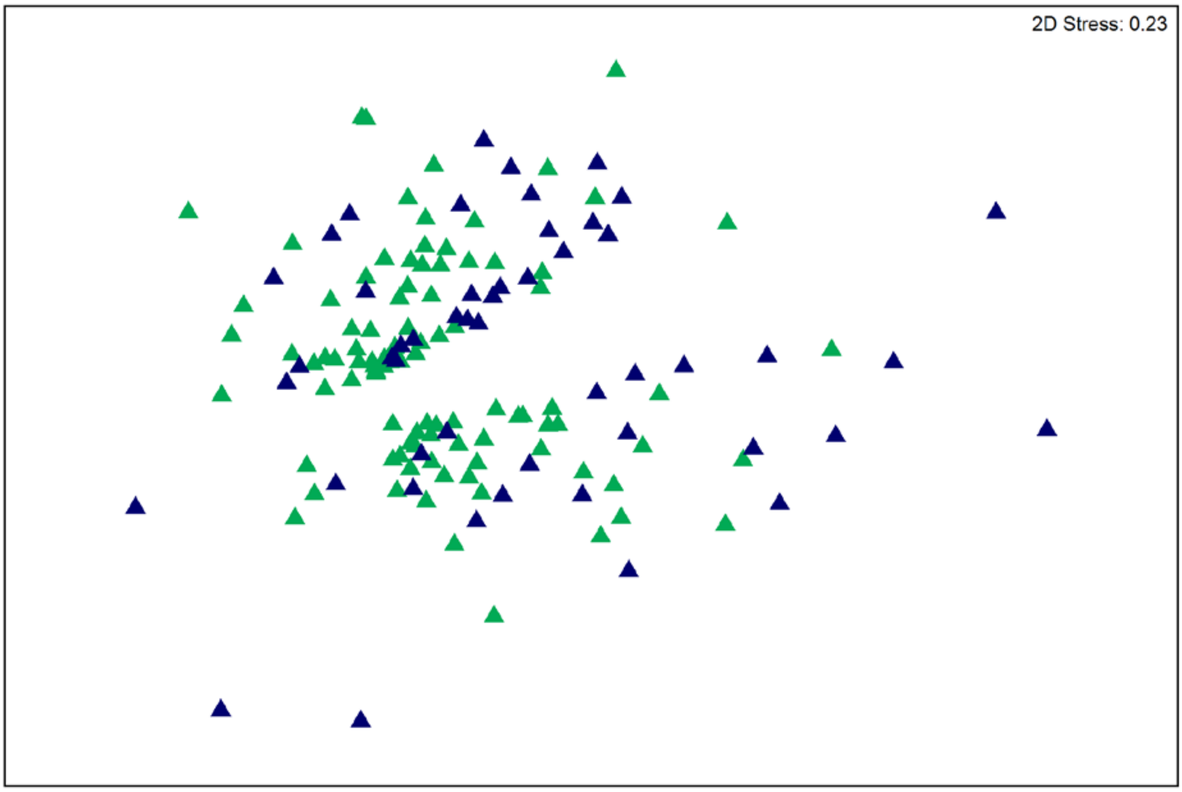


Figure 3.16 A two-dimensional MDS plot of the development sub-profile of educator perspectives on CT (n=146).Educators are coloured by the education setting in which they teach. The green triangles represent the high school educators and the dark blue triangles represent the tertiary educators.

Table 3.10 Summary of the contributions of the items for each question, to factors with significant group differences in the development sub-profile.

respondent type	sig. level^		Question Set 2 (15 items)	Question Set 3 (3 items)	Question Set 4 (18 items)
Science vs History	***	% contribution to group differences	23.65	11.05	65.33
		min - max %	0.92 - 2.11	0.97 - 5.07	2.22 - 5.76
High school vs Tertiary	***	% contribution to group differences	23.53	10.97	65.5
		min - max %	1.01 - 2.02	1.45 - 4.89	2.44 - 4.36
Science: High school vs Tertiary	***	% contribution to group differences	23.85	11.05	65.13
		min - max %	1.03 - 2.03	1.78 - 4.83	2.6 - 4.59

^ significance level based on PERMANOVA results (refer to Table 4.5);
*denotes p<0.05, **denotes p<0.01, ***denotes p<0.001

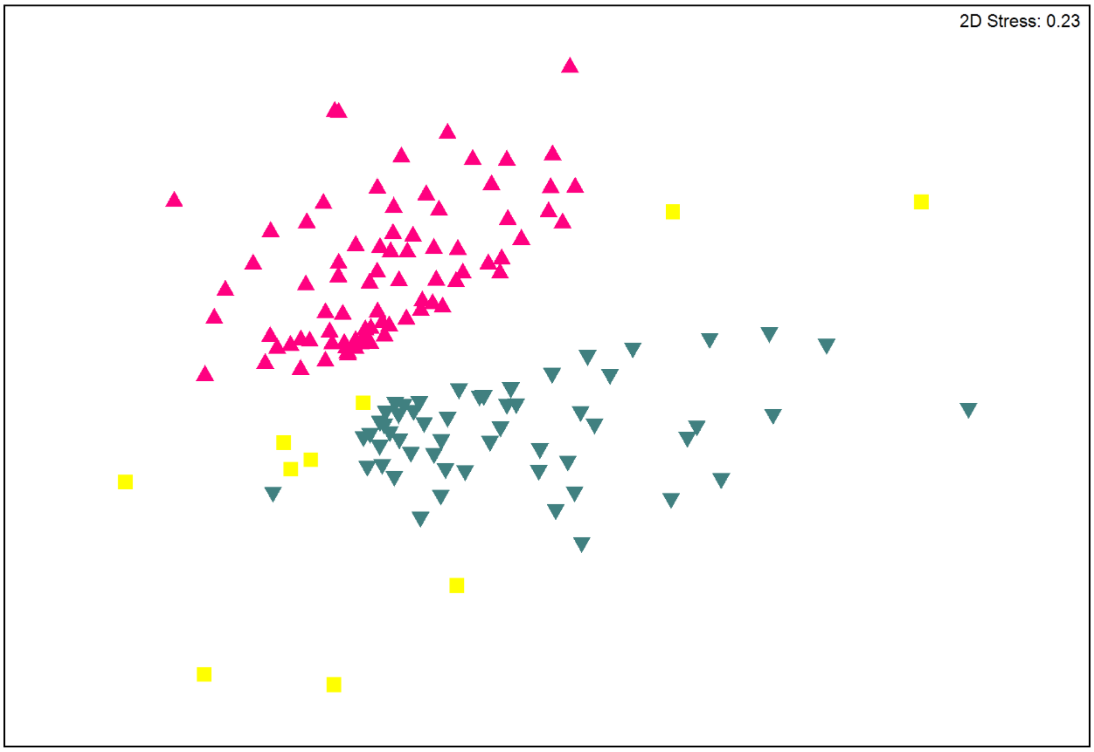


Figure 3.17 A two-dimensional MDS plot of the development sub-profile of educator perspectives on CT (n=146).Educators are coloured by their response to Question 3. The pink triangles represent educators who use an embedded approach. The teal triangles represent the educators who use an explicit development approach. The yellow squares represent the educators who indicated they had no specific plan.

Table 3.11 Summary of group dissimilarity analysis from the SIMPER approach for demographic factors identified as significant in the development sub-profile. Displayed in rank order of contributions to the differences by education setting (i.e. between high school and tertiary educators).

Item code	Item theme	SIMPER rank (% contribution of item to difference between groups; group with higher mean score)					
		Education setting (HS vs TER)		Science: Education setting (Sc HS v Sc TER)		Discipline (Sc v His)	
D_3b	**approaches embedded across the curriculum**	**1**	**(4.89%; HS)**	**1**	**(4.83%; Sc HS)**	2	(5.07%; Sc)
D_3a	approaches involving explicit instruction	2	(4.63%; HS)	3	(4.44%; Sc HS)	3	(5.01%; His)
D_4c	concept mapping effectiveness	3	(4.36%; HS)	2	(4.59%; Sc HS)	13	(3.67%; His)
D_4k	inductive reasoning activities	4	(4.28%; HS)	5	(4.28%; Sc HS)	5	(4.27%; Sc)
D_4r	argument analysis effectiveness	5	(4.27%; HS)	4	(4.34%; Sc HS)	6	(4.13%; His)
D_4i	socratic questioning effectiveness	6	(4.23%; HS)	9	(4.1%; Sc HS)	4	(4.56%; His)
D_4l	deductive reasoning effectiveness	7	(4.1%; HS)	10	(3.94%; Sc HS)	8	(3.99%; Sc)
D_4d	logic models effectiveness	8	(4.09%; HS)	6	(4.28%; Sc HS)	9	(3.94%; Sc)
D_4f	debating effectiveness	9	(4.09%; HS)	7	(4.24%; Sc HS)	10	(3.85%; His)
D_4m	constucting critiques effectiveness	10	(4.06%; HS)	8	(4.17%; Sc HS)	7	(4%; His)
D_4h	**scientific method effectiveness**	11	(3.79%; HS)	18	(2.75%; Sc HS)	**1**	**(5.76%; Sc)**
D_4e	context dependant sets effectiveness	12	(3.68%; HS)	11	(3.7%; Sc HS)	12	(3.7%; Sc)
D_4j	peer assessment effectiveness	13	(3.66%; HS)	12	(3.57%; Sc HS)	14	(3.58%; His)
D_4g	flipped classroom effectiveness	14	(3.57%; HS)	13	(3.53%; Sc HS)	11	(3.75%; His)
D_4p	case studies effectiveness	15	(3.42%; HS)	14	(3.5%; Sc HS)	15	(3.16%; His)
D_4q	writing tasks effectiveness	16	(2.98%; TER)	16	(3.06%; Sc TER)	16	(2.94%; His)
D_4b	collaborative learning effectiveness	17	(2.96%; HS)	15	(3.09%; Sc HS)	18	(2.63%; His)
D_4o	oral presentation effectiveness	18	(2.83%; HS)	17	(2.76%; Sc HS)	17	(2.77%; His)
D_4a	problem based learning effectiveness	19	(2.69%; HS)	20	(2.6%; Sc HS)	19	(2.41%; Sc)
D_4n	class discussion effectiveness	20	(2.44%; HS)	19	(2.63%; Sc HS)	20	(2.22%; His)
D_2b	citizenship vs CT	21	(2.02%; TER)	21	(2.03%; Sc TER)	24	(1.93%; His)
D_2g	intercultural understanding vs CT	22	(1.9%; TER)	23	(1.84%; Sc TER)	22	(2.04%; Sc)
D_2e	ethical conduct vs CT	23	(1.88%; TER)	22	(1.94%; Sc TER)	25	(1.81%; Sc)
D_2j	numeracy vs CT	24	(1.78%; TER)	30	(1.61%; Sc TER)	21	(2.11%; His)
D_2h	knowledge vs CT	25	(1.77%; TER)	28	(1.71%; Sc TER)	23	(1.95%; Sc)
D_2i	literacy vs CT	26	(1.77%; TER)	26	(1.78%; Sc TER)	26	(1.76%; Sc)
D_2n	teamwork vs CT	27	(1.72%; TER)	24	(1.79%; Sc TER)	27	(1.72%; His)
D_2f	ICT vs CT	28	(1.65%; TER)	29	(1.69%; Sc TER)	28	(1.66%; His)
D_2o	understanding vs CT	29	(1.61%; HS)	27	(1.73%; Sc HS)	29	(1.52%; His)
D_2k	planning vs CT	30	(1.53%; TER)	32	(1.54%; Sc TER)	30	(1.5%; His)
D_2c	communication vs CT	31	(1.47%; TER)	31	(1.57%; Sc TER)	31	(1.35%; His)
D_3c	no plan	32	(1.45%; TER)	25	(1.78%; Sc TER)	35	(0.97%; Sc)
D_2d	creativity vs CT	33	(1.31%; TER)	33	(1.33%; Sc TER)	32	(1.34%; His)
D_2l	problem solving vs CT	34	(1.1%; TER)	34	(1.15%; Sc TER)	34	(1.01%; His)
D_2a	analysis vs CT	35	(1.01%; TER)	35	(1.11%; Sc TER)	36	(0.92%; His)
D_2m	reasoning vs CT	36	(1.01%; TER)	36	(1.03%; Sc TER)	33	(1.03%; His)

3.4.3 Concerning understanding about CT assessment

Educators responded to three survey questions (two quantitative and one qualitative) relating to their understanding of the assessment of CT. Findings from the qualitative question are presented as a word cloud. Quantitative questions were explored using the PERMANOVA and SIMPER routines at a question level, and as a theoretical set, where the items from two quantitative assessment questions were incorporated into the assessment sub-profile. The results, contributions of specific question items to group differences and trends associated with each question and the assessment sub-profile (including by any demographic factors) are explained below.

Findings related to the hypothesis testing undertaken on the quantitative questions and assessment sub-profile using the PERMANOVA routine.

When specifically evaluating the CT assessment sub-profile, education setting was the only the demographic factor that was determined to be significantly different at the 0.01 level. This was true of both the entire response set (Table 3.12: pseudo-F = 8.98, p-value = 0.0001), and within the science cohort (Table 3.12: pseudo-F = 7.89, p-value = 0.0001). This same trend of significance was also present for Question 5 (assessment tools), but not Question 6 (assessment perspectives), in the individual question level analysis (Table 3.12). There were no significant interaction effects for any of the demographic factors. Again, the SIMPER routine and descriptive techniques were used to explore reasons for the differences in perspectives by respondent's education setting. These findings are presented next.

Table 3.12 Summary of findings for *a priori* hypothesis tests undertaken within the assessment question-set, using PERMANOVA to explore the effects of key demographic factors on educator perceptions about CT.

	PERMANOVA results for the assessment sub-profile						PERMANOVA results for the assessment sub-profile		
	Q5 assessment tools			Q6 assessment perspectives					
factor	pseudo-F	p-value	df	pseudo-F	p-value	df	pseudo-F	p-value	df
discipline	1.95	0.1014	1	2.97	0.0159	1	2.03	0.0776	1
education setting	**9.74**	**0.0001**	**1**	1.18	0.3087	1	**8.98**	**0.0001**	**1**
gender	1.87	0.1154	1	1.83	0.1115	1	1.86	0.1112	1
state	1.32	0.1430	6	1.00	0.4534	6	1.29	0.1512	6
teaching experience	1.92	0.0592	2	1.06	0.3882	2	1.85	0.0603	2
discipline vs education setting	1.45	0.2203	1	0.13	0.9669	1	1.08	0.3794	1
discipline vs gender	1.76	0.1508	1	0.10	0.9719	1	1.62	0.1607	1
discipline vs state*	1.26	0.2145	4	0.72	0.7904	4	1.21	0.2524	4
discipline vs teaching experience	0.74	0.6701	2	1.60	0.1048	2	0.83	0.5986	2
education setting vs gender	0.30	0.8482	1	0.67	0.6337	1	0.32	0.8698	1
education setting vs state*	1.34	0.1750	4	1.31	0.1757	4	1.40	0.1260	4
education setting vs teaching experience	0.90	0.5337	2	1.84	0.0604	2	0.97	0.4630	2
gender vs state*	0.76	0.7165	6	1.04	0.4110	6	0.64	0.9203	6
gender vs teaching experience	1.05	0.4035	2	0.30	0.9721	2	0.93	0.5088	2
state* vs teaching experience	1.00	0.4705	8	0.99	0.4907	8	1.01	0.4615	8
science - education setting	**8.63**	**0.0001**	**1**	1.03	0.3877	1	**7.89**	**0.0001**	**1**
science - gender	1.55	0.1998	1	1.30	0.2578	1	1.55	0.1784	1
science - state	1.52	0.0531	6	1.11	0.3114	6	1.48	0.0548	6
science - teaching experience	1.39	0.2053	2	0.86	0.5620	2	1.35	0.2115	2

*pairwise tests incorporating state as a factor excluded NT and ACT respondents from analysis due to low sample size.

Trends for Question 5

Question 5 asked educators to indicate the types of assessment approaches they thought would be suitable for assessing CT in their educational context. All but one of the assessment strategies were selected by at least 50% of the respondents as a strategy that educators would use (Figure 3.18). The most popular assessment strategies were extended response items and discipline-specific CT tests, where over 80% of educators indicated they would use these approaches to assess CT (Figure 3.18). Multiple ranking items were the least chosen assessment approach and was also skipped by the greatest number of respondents.

To seek patterns within the responses (in other words, explore whether it was the same educators who showed a preference for a particular tool development style), a similarity matrix and a Principal Coordinates Analysis (PCoA) ordination were generated for Question 5 responses. Examination of the PCoA with vector overlays for each discipline response set (Figure 3.19 A-B and Figure 3.20 A-B), highlighted that trends in educator responses seemed to relate more to the design of the tool (self-developed or externally developed), than a trend towards any particular tool type (which is the hypothesis test that was undertaken). It also revealed that there was less variation in the science educator's choices, whose vectors point in three clear directions (each corresponding to a particular tool development approach). Whereas the clusters of vectors based on the history educator choices display a mix of tool development approaches (see Figure 3.20 A-B).

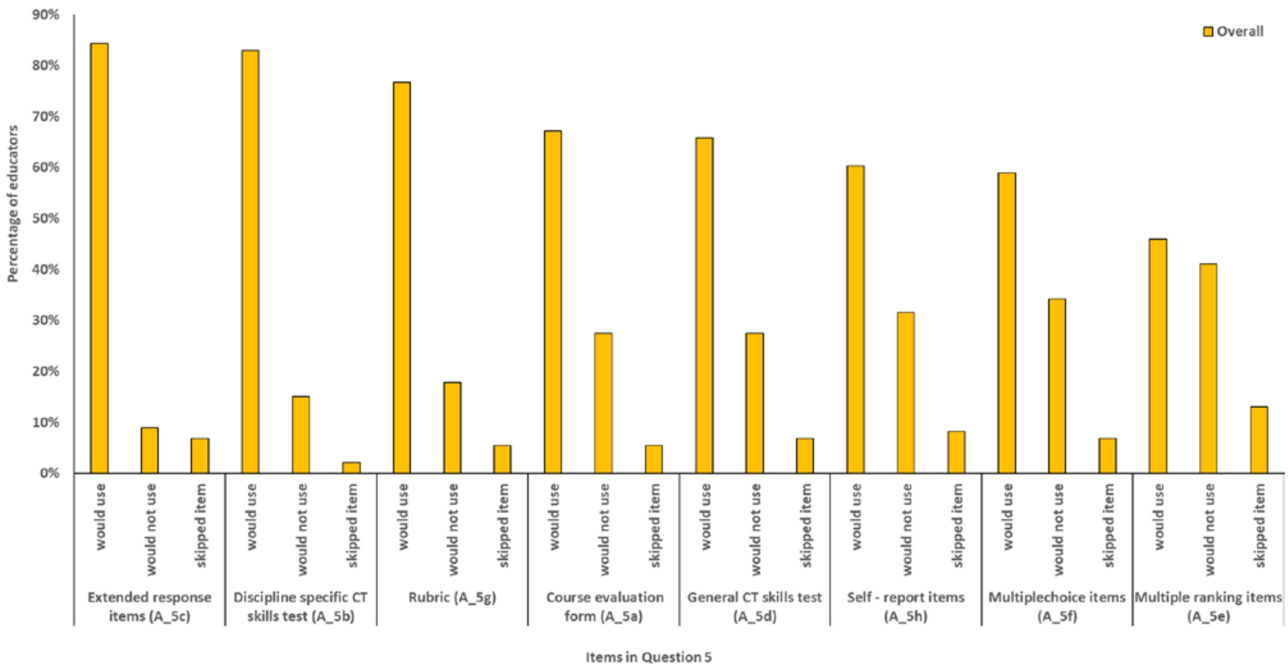
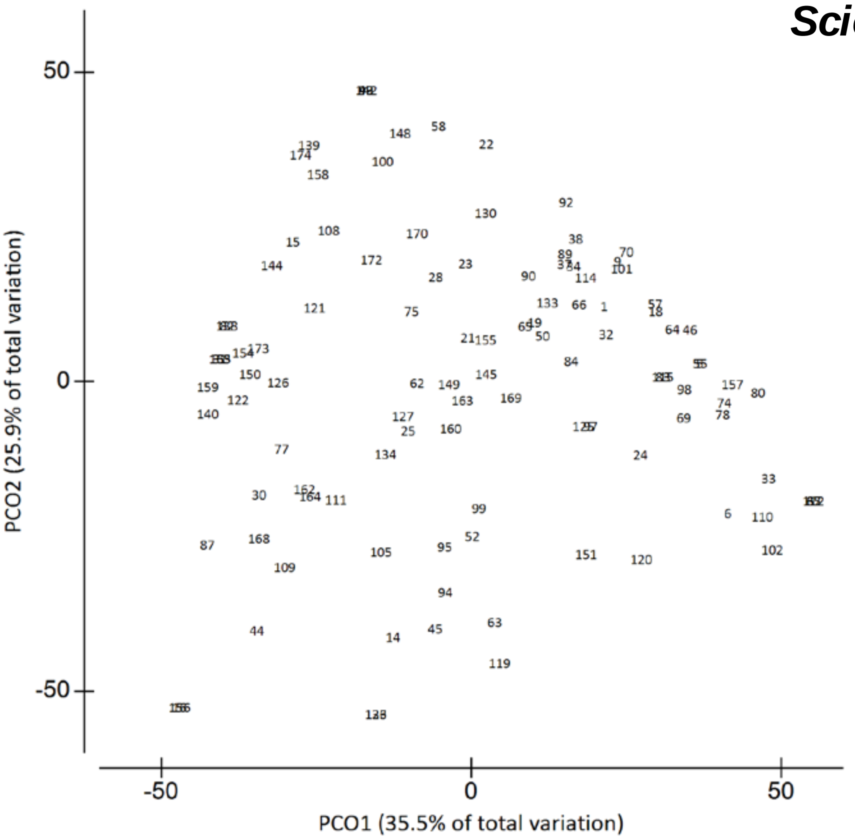


Figure 3.18 Summary of all educator responses to Question 5 (n=146): Which of the following strategies do you think you would use to assess CT at your educational institution? Assessment options are rank ordered from highest to lowest reported use.

(A) ***Science***



(B)

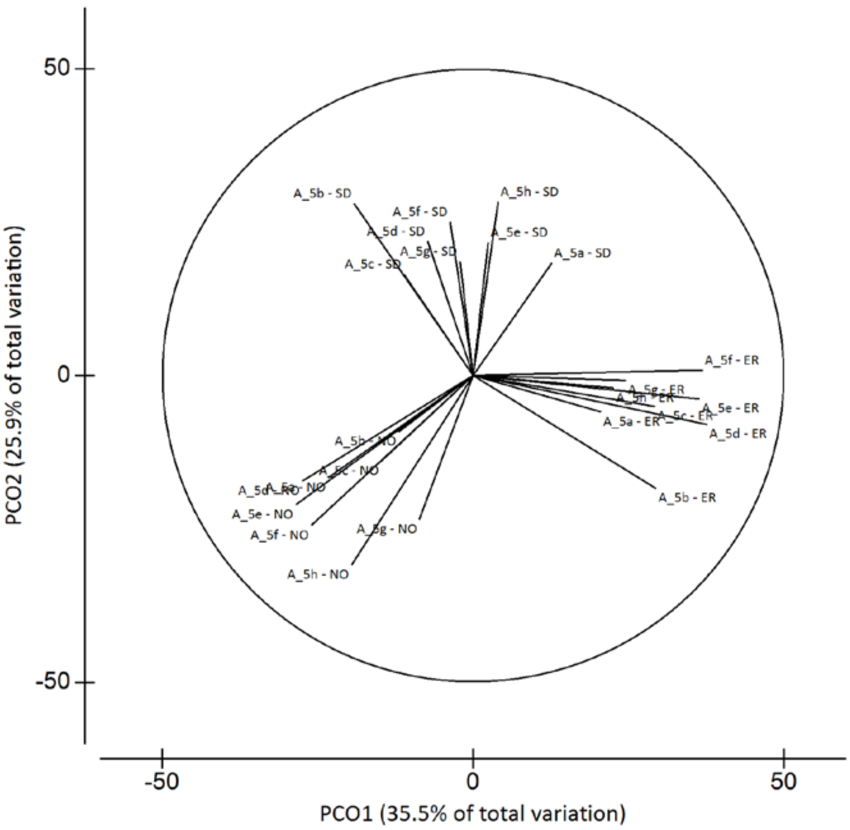


Figure 3.19 A-B (A) Two-dimensional PCoA of science educator responses to Question 5 (assessment tool use) (n=110). (B) shows vectors of assessment tool items overlayed on the corresponding ordination (above), with each vector line indicating the extent of correlation of the responses to items, in addition to collectively highlighting what respondents positioned in that part of the data cloud answered similarly.

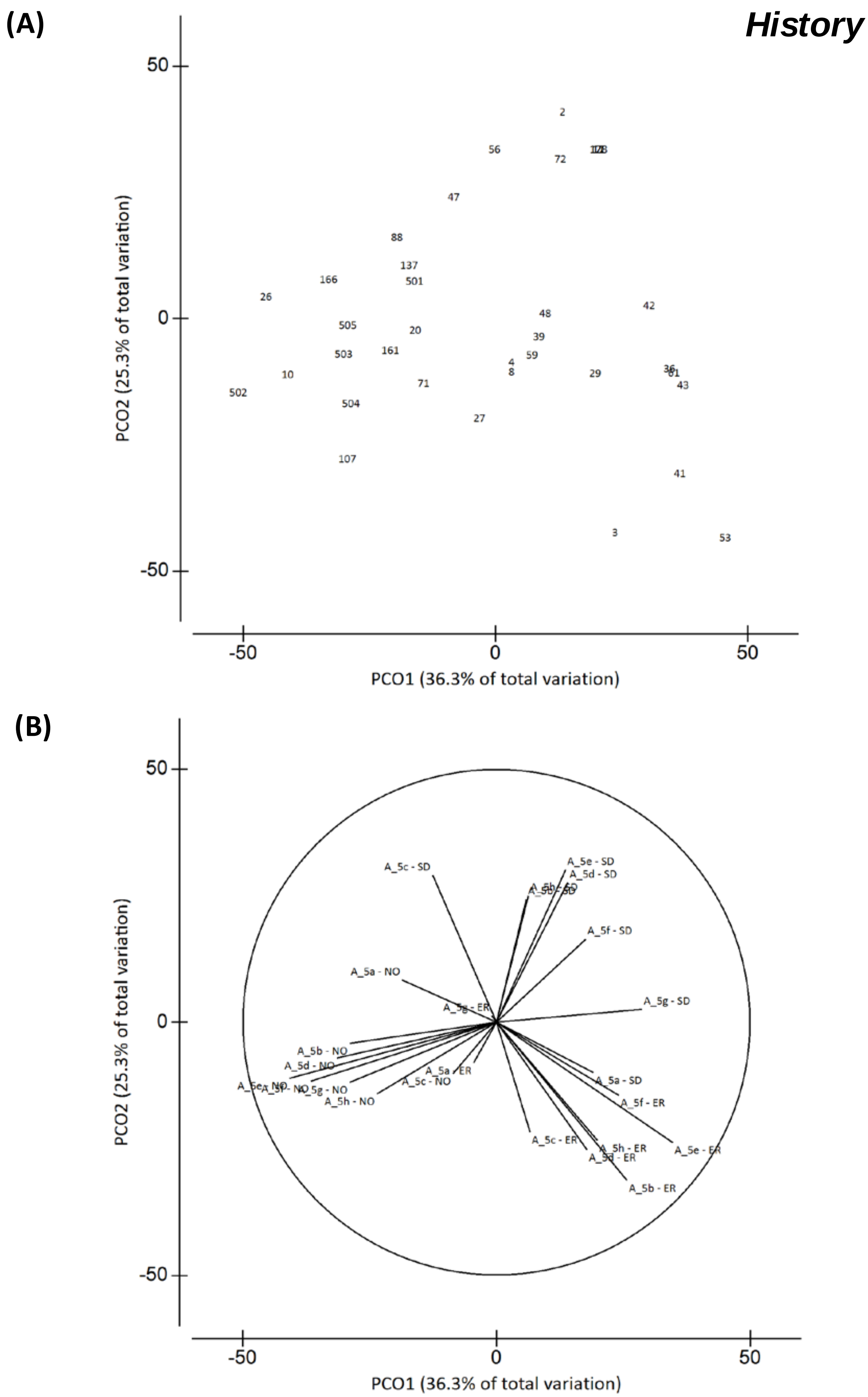


Figure 3.20 A-B (A) Two-dimensional PCoA of history educator responses to Question 5 (assessment tool use) (n=36). (B) shows vectors of assessment tool items overlayed on the corresponding ordination (above), with each vector line indicating the extent of correlation of the responses to items, in addition to collectively highlighting what respondents positioned in that part of the data cloud answered similarly.

Responses were again explored by demographic factors to determine whether an educator's context influenced the assessment strategies they would use in their setting; however, exploration by the majority of the factors revealed few significant differences. For example, science and history educators responded similarly across the item types, where only slightly more science educators than history educators indicated they would use approaches such as multiple ranking items (A_5e), discipline-specific CT test (A_5b), and general CT tests (A_5d) (see Table 3.13). But when exploring whether perspectives varied by education setting, it was identified that there were significances differences in the way high school and tertiary educators responded across items (Table 3.12: PERMANOVA for education setting: pseudo-F = 9.74, p-value = 0.0001; and PERMAVONVA for science: education setting: pseudo-F = 8.63, p-value = 0.0001). SIMPER was again used to explain the items contributing the most to group differences, and this revealed that in general, similar items were contributing to differences by each significant demographic factor (Table 3.14); however, the extent of the contribution per item varied (resulting in a different rank order).

For the demographic factor education setting (all respondents), SIMPER analysis revealed that the two best discerning items for assessing CT were A_5a - SD and A_5b – SD (see Table 3.14) which address the use of self-developed course evaluation forms (A_5a - SD) and self-developed discipline-specific CT tests (A_5b - SD). High school educators indicated they would be more likely to use self-developed course evaluation forms to assess CT than the tertiary educators would, whereas tertiary educators were more likely to use a self-developed discipline-specific CT test than high school educators (Table 3.13). However, there were many other points of difference, with many items contributing similarly to group differences (Table 3.14).

In the science cohort, the use of self-developed discipline-specific CT tests (A_5b - SD) was again important for discerning group differences, as were educator perspectives on the use of externally sourced general CT tests (A_5d - ER) (see Table 3.14). The tertiary science educators indicated they would be more likely to use a self-developed discipline-specific CT test than the high school science educators, whereas the high school science educators indicated they would be less likely to use general CT tests to assess CT than the tertiary science educators (Table 3.13).

Generally, these patterns show that there are both some education setting and discipline-influenced choices that high school and tertiary educators make in relation to CT assessment.

Table 3.13 Proportions of educators who chose each option for Question 5 items 5a, 5b, 5c, and 5d for demographic factors: discipline and education setting (overall and within the science cohort). Proportions for all Question 5 items are found in Appendix C: Table C.6.

Assessment approach		Demographic factor					
		Education setting		Science: Education setting		Discipline	
Item theme and code	Options	Tertiary (n=50)	High school (n=96)	Tertiary (n=41)	High school (n=69)	Science (n=110)	History (n=36)
Course evaluation form (A_5a)	Self - developed item	34%	59%	34%	58%	49%	56%
	External resource	18%	25%	15%	28%	23%	22%
	Would not use	40%	21%	41%	22%	29%	22%
	Skipped	10%	3%	12%	1%	5%	6%
Discipline specific CT skills test (A_5b)	Self - developed item	60%	40%	61%	36%	45%	50%
	External resource	30%	56%	37%	59%	51%	36%
	Would not use	18%	14%	12%	13%	13%	22%
	Skipped	2%	2%	2%	1%	2%	3%
General CT skills test (A_5d)	Self - developed item	26%	25%	29%	23%	25%	25%
	External resource	32%	56%	34%	61%	51%	39%
	Would not use	40%	21%	34%	20%	25%	33%
	Skipped	8%	6%	10%	6%	7%	6%
Multiple ranking items (A_5e)	Self - developed item	18%	20%	20%	20%	20%	17%
	External resource	14%	41%	17%	45%	35%	22%
	Would not use	52%	35%	44%	33%	37%	53%
	Skipped	18%	10%	22%	10%	15%	8%

Table 3.14 SIMPER dissimilarity results for Question 5 for demographic factors identified as significantly different through the PERMANOVA routine. Order of items displayed to indicate the highest to lowest contribution to group differences by the demographic factor education setting.

Item description	Item code	SIMPER rank (% contribution of item to difference between groups; group with higher mean score)			
		Education setting (HS vs TER)		Science: Education setting (Sc HS v Sc TER)	
self-developed course evaluation form	**A_5a - SD**	**1**	**(5.2%; HS)**	4	(5.11%; Sc TER)
self-developed discipline specific CT test	A_5b - SD	2	(5.19%; TER)	2	(5.23%; Sc TER)
externally sourced discipline specific CT	A_5b - ER	3	(5.15%; HS)	3	(5.15%; Sc HS)
externally sourced general CT test	**A_5d - ER**	4	(5.14%; HS)	**1**	**(5.23%; Sc HS)**
self developed rubric	A_5g - SD	5	(5.04%; HS)	7	(4.93%; Sc HS)
would not use multiple ranking items	A_5e - NO	6	(4.94%; TER)	10	(4.62%; Sc TER)
self-developed self-report items	A_5h - SD	7	(4.92%; HS)	8	(4.81%; Sc HS)
self-developed extended response	A_5c - SD	8	(4.79%; HS)	6	(4.95%; Sc HS)
would not use multiple choice items	A_5f - NO	9	(4.56%; TER)	15	(4.13%; Sc TER)
would not use self-report items	A_5h - NO	10	(4.5%; TER)	12	(4.37%; Sc TER)
externally sourced extended response	A_5c - ER	11	(4.49%; HS)	5	(5.02%; Sc HS)
externally sourced multiple choice item	A_5f - ER	12	(4.44%; HS)	9	(4.81%; Sc HS)
self-developed multiple choice items	A_5f - SD	13	(4.27%; TER)	13	(4.31%; Sc TER)
would not use course evalaution form	A_5a - NO	14	(4.27%; TER)	14	(4.29%; Sc TER)
would not use general CT test	A_5d - NO	15	(4.25%; TER)	16	(3.87%; Sc TER)
externally sourced multiple-ranking	A_5e - ER	16	(4.17%; HS)	11	(4.46%; Sc HS)
self-developed general CT skills tests	A_5d - SD	17	(3.78%; TER)	17	(3.82%; Sc TER)
would not use rubric	A_5g - NO	18	(3.55%; TER)	19	(3.26%; Sc TER)
externally sourced rubric	A_5g - ER	19	(3.45%; HS)	18	(3.66%; Sc HS)
externally sourced course evaluation	A_5a - ER	20	(3.24%; HS)	20	(3.24%; Sc HS)
self-developed multiple ranking items	A_5e - SD	21	(2.87%; HS)	22	(2.91%; Sc HS)
externally sourced self-report items	A_5h - ER	22	(2.74%; HS)	21	(2.97%; Sc HS)
would not use discipline specific CT test	A_5b - NO	23	(2.57%; TER)	24	(2.12%; Sc HS)
would not use extended response items	A_5c - NO	24	(2.48%; TER)	23	(2.72%; Sc TER)

Trends for Question 6

For Question 6, which addresses assessment perspectives, educators were asked to allocate points to statements to indicate which reflected their current approach to assessing CT. In general, most educators indicated they look for specific evidence of CT embedded in students work (theme of item A_6d). As shown in Figure 3.21 A and B, the weighting of points to the most frequently chosen statement (A_6d: look for specific evidence) was similar to the weighting of points allocated to the alternative perspective which had been chosen less frequently (do not look for specific evidence: i.e. themes of items A_6b and A_6c). There was slightly more variation in the point allocations to statements concerning specific ways the educators assess CT (i.e. Figure 3.21B: statements A_6e through A_6h).

Exploration of the results by discipline revealed that statement selections were generally similar. Interestingly, less than 6% of educators indicated that they had trouble assessing CT. Yet 39% of science educators and 23% of history educators indicated they did not look for specific evidence to assess CT (sum of the frequency of items 6b and 6c, Figure 3.22A). This same trend was also displayed tertiary and high school educators (Figure 3.23 A-B) and in the results from the science only cohort (Figure 3.24 A-B). However, there were two items, A_6a and A_6c, where the tertiary educators proportionally spent more points on their statement choices than the high school educators who also included this statement in their perspective (Figure 3.23 A-B and 3.24 A-B). This indicated a stronger alignment to these tertiary educators' perceptions about this aspect of CT assessment (not looking for specific evidence but knowing students are developing CT as their marks improve) than the for high school educators who also chose to include this statement in their perspective.

For this question, there were no significant differences by any of the demographic factors (refer back to Table 3.12).

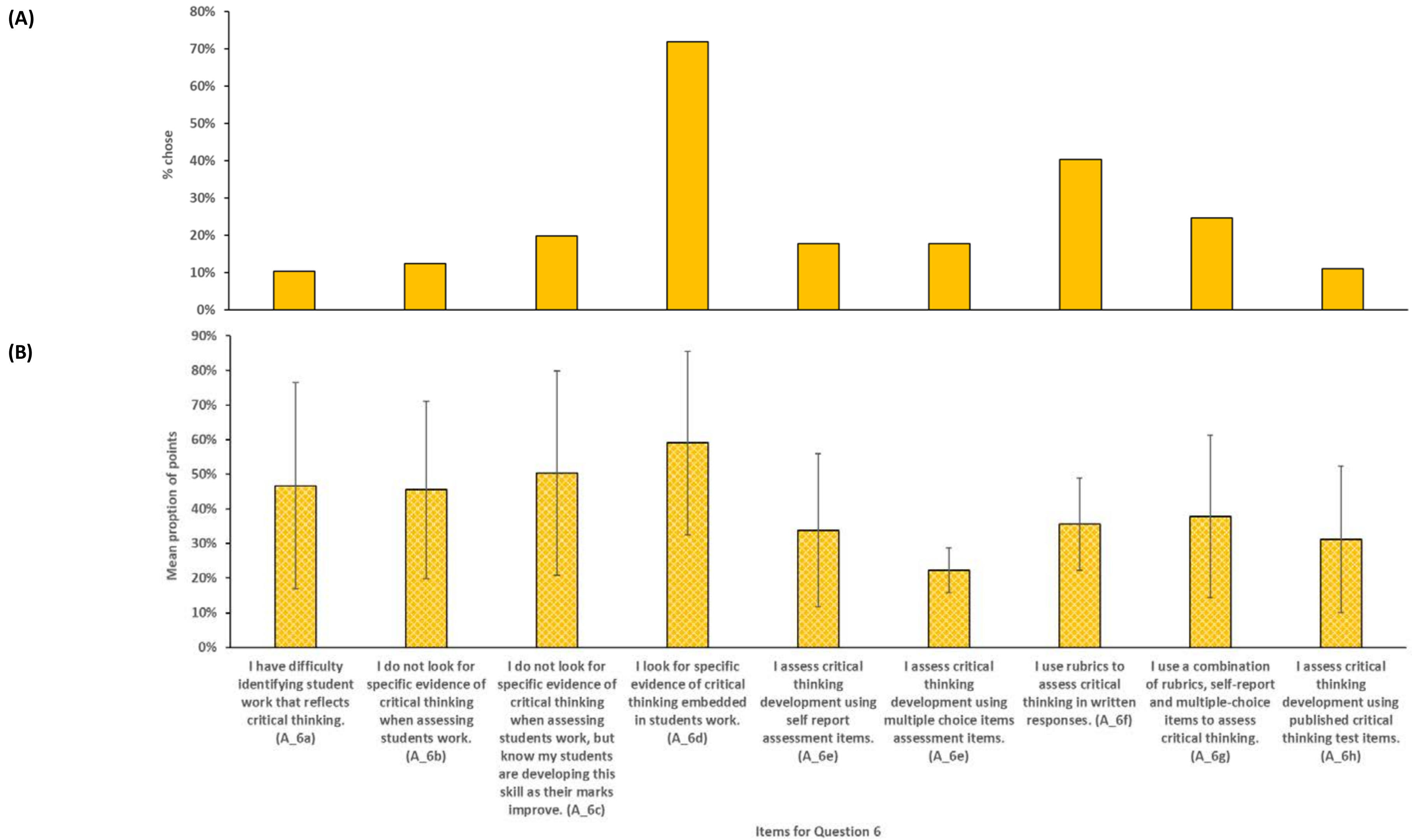


Figure 3.21 A-B Distribution of educator choices (A) and point allocations (B) to Question 6, comparing the perspectives on assessing CT (n=146). Solid bars represent the percentage respondents that assigned at least one point to the statement. The patterned bar represents the mean proportion (%) of points allocated to the statement by educators who chose that statement, with the higher the number of points allocated indicating the stronger the statement aligns with their viewpoint. Error bars represent ± 1 SD.

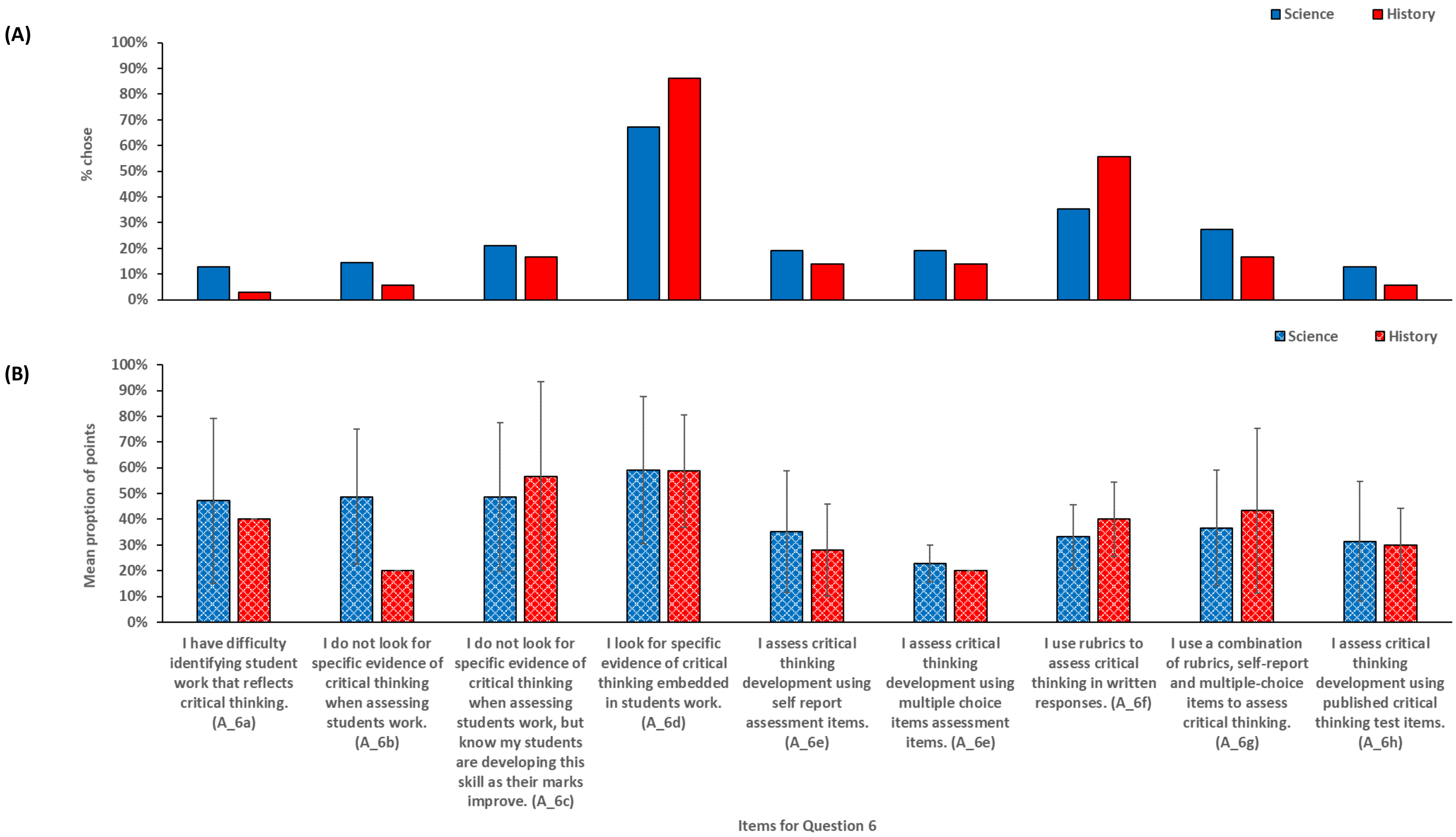


Figure 3.22 A–B Distribution of science (n=110) and history (n=36) educator choices (A) and point allocations (B) to Question 6, comparing the perspectives on assessing CT. Solid bars represent the percentage respondents that assigned at least one point to the statement. The criss-cross patterned bar represents the mean proportion (%) of points allocated to the statement by each educator who chose that statement from that group, with the higher the number of points allocated indicating the stronger the statement aligns with their viewpoint. Error bars represent ± 1 SD.

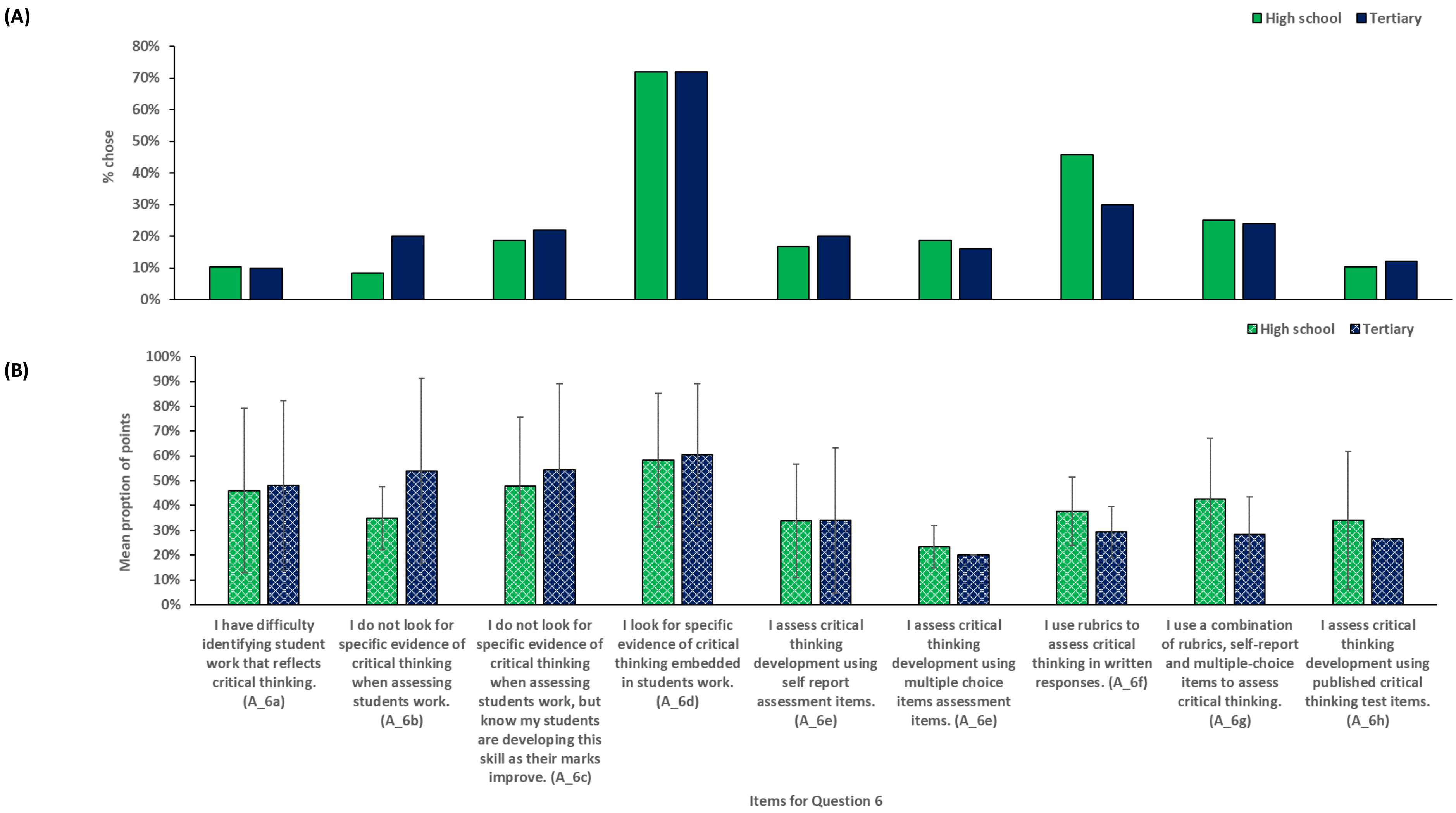


Figure 3.23 A-B Distribution of tertiary (n=50) and high school (n=96) educator choices (A) and point allocations (B) to Question 6, comparing the perspectives on assessing CT.Solid bars represent the percentage respondents that assigned at least one point to the statement. The criss-cross patterned bar represents the mean proportion (%) of points allocated to the statement by each educator who chose that statement from that group, with the higher the number of points allocated indicating the stronger the statement aligns with their viewpoint. Error bars represent ± 1 SD.

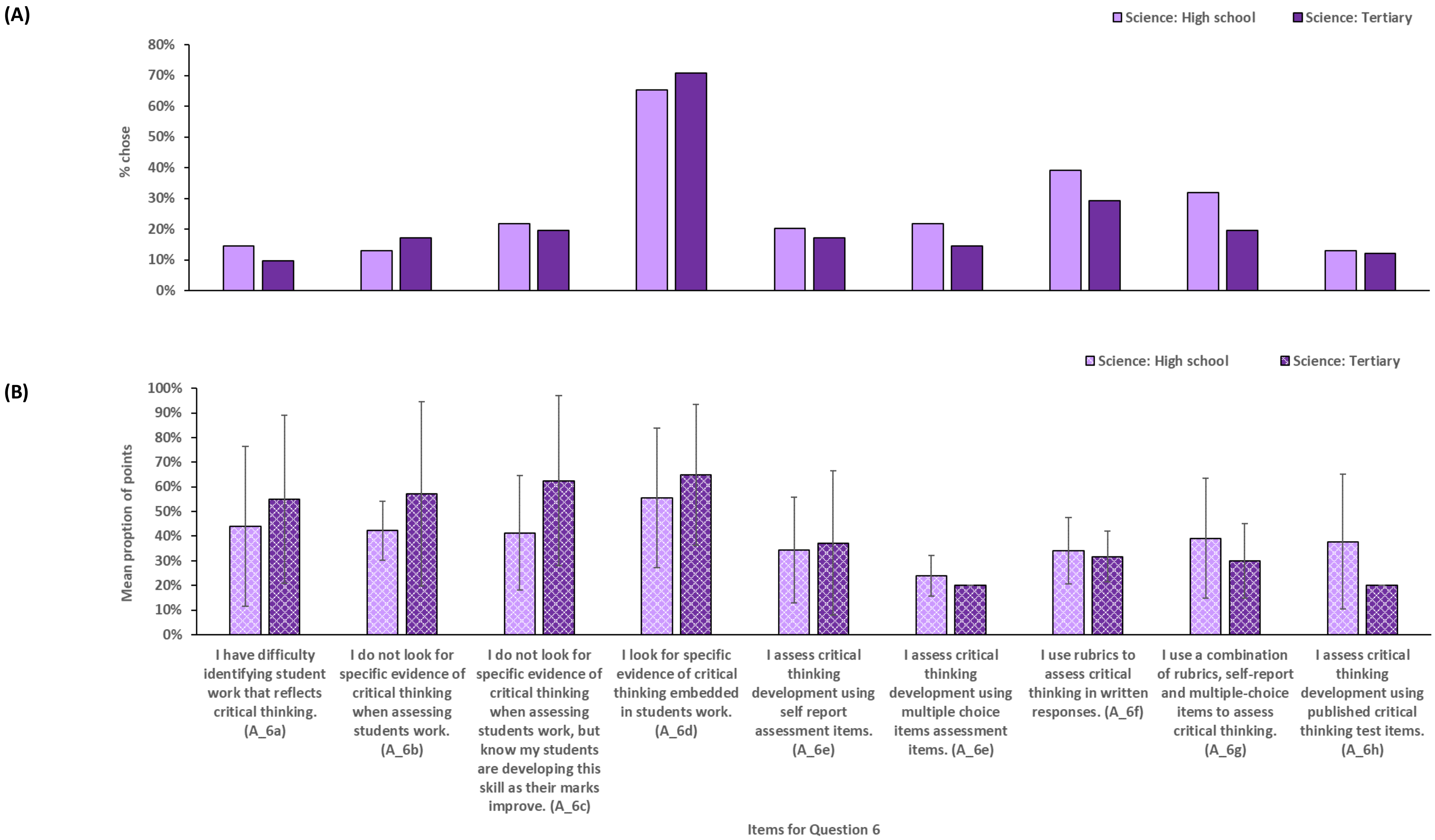


Figure 3.24 A–B Distribution of tertiary science (n=41) and high school science (n=69) educator choices (A) and mean point allocations (B) to Question 6, comparing the perspectives on assessing CT. Solid bars represent the percentage respondents that assigned at least one point to the statement. The criss-cross patterned bar represents the mean proportion (percentage) of points allocated to the statement by each educator who chose that statement from that group, with the higher the number of points allocated indicating the stronger the statement aligns with their viewpoint. Error bars represent ± 1 SD.

Trends for 6a

For Question 6a, educators were asked if there were any other strategies they would implement to assess CT development. Thirty-six educators (approximately 25%) provided an example. Four main themes emerged from this question (see Figure 3.25). The two most common were verbal assessments (formal questioning, informal questioning, and discussions), and using samples of classwork (with CT embedded) to assess CT. The two other main themes were research tasks (including experimental design, question development, experiment evaluation) and career-resembling tasks were also a common included idea. Two educators specifically noted that the majority of assessment tools options in Question 6 were written, and several others included terms suggesting informal rather than formal assessments. One educator did not provide an example, but instead responded that assessing CT in isolation was not something they had thought about before, but would now. Collectively, these indicated that CT assessment is not always a formal priority.

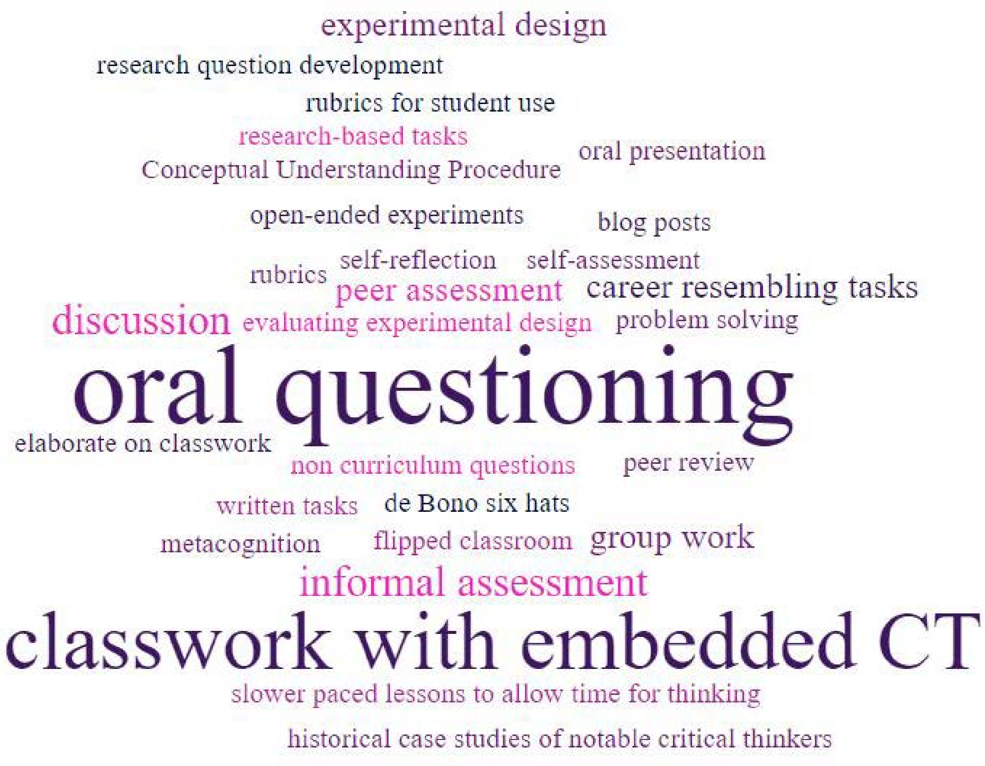


Figure 3.25 Core themes and terms included in educator responses to Question 6a: Are there any other strategies you would implement to assess CT development?

Trends for assessment sub-profile

For the assessment sub-profile, there are no clear clusters when examining the ordination (Figure 3.26); however, there is a greater density of respondents shifted to the left of the data cloud. Colouring respondents in this data cloud respondent's education setting reveals that this left shift is mostly comprised of high school educators (Figure 3.27). This concentration of educators within the data cloud is also evident in the science cohort ordination (Figure 3.28); however, the shift is to the right. In both cases, there is a greater spread, and thus more variability, in the way that the tertiary educators indicated they assess CT. PERMANOVA analysis confirmed that there were significant differences in perspectives by the demographic factor education setting (refer back to Table 3.12: all respondents: pseudo-F = 9.74, p-value = 0.0001; science respondents: pseudo-F = 7.89, p-value = 0.0001). None of the other demographic factors or interactions were determined to be significantly different.

The SIMPER routine was again used to explore the differences in perspectives by education setting and science: education setting. This revealed that Question 5 contributed more to group differences than items from Question 6 (Table 3.15). In addition, the same four items contributed the most to group differences for both education-setting factors (Table 3.16). However, the rank order of the items differed, with selections around the use of self-developed evaluation form (theme of item A_5a-SD) being more distinguishing between the high school and tertiary educators. Whereas selections on the use of externally sourced general CT test (theme of item A_5d-ER) being more distinguishing within the science cohort at the different education setting levels.

Overall, exploration of the differences in the assessment sub-profile revealed that education setting influences the choices that educators make about CT assessment.

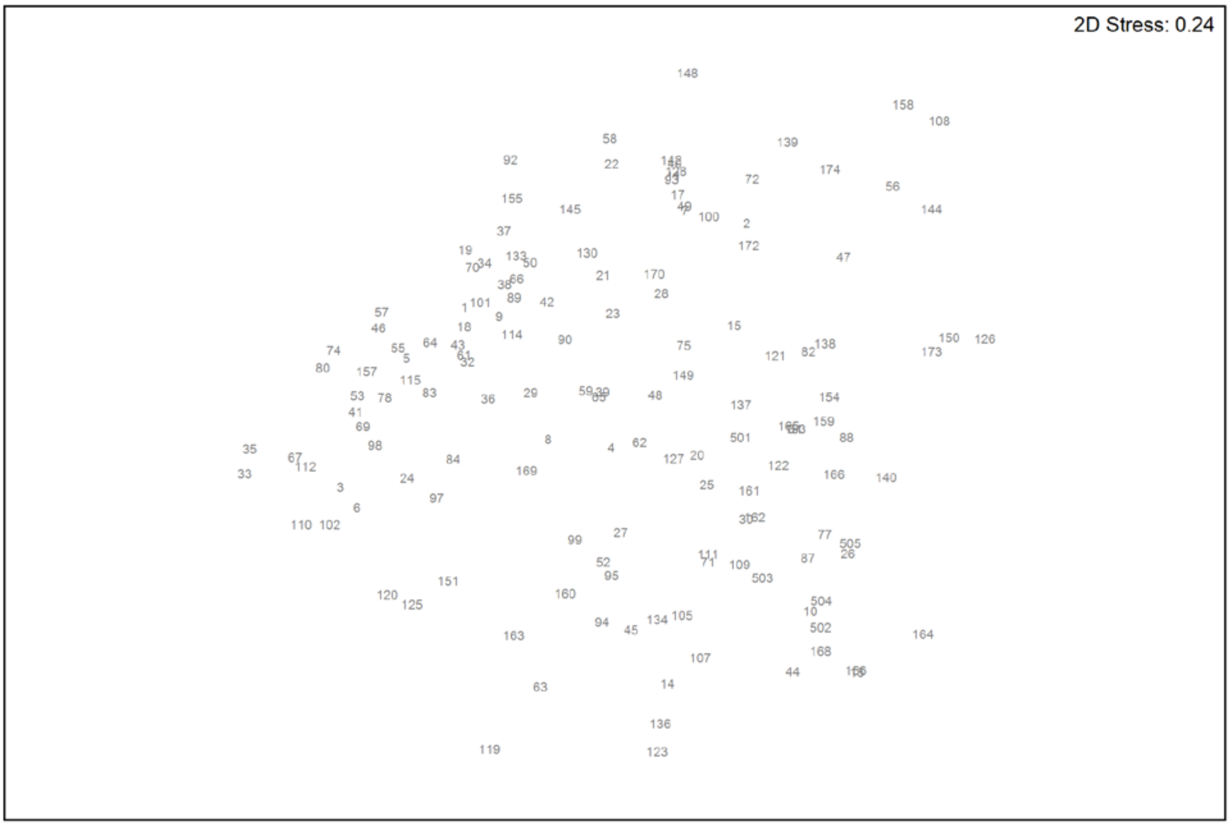


Figure 3.26 A two-dimensional MDS plot of the assessment sub-profile of educator perspectives on CT (n=146).

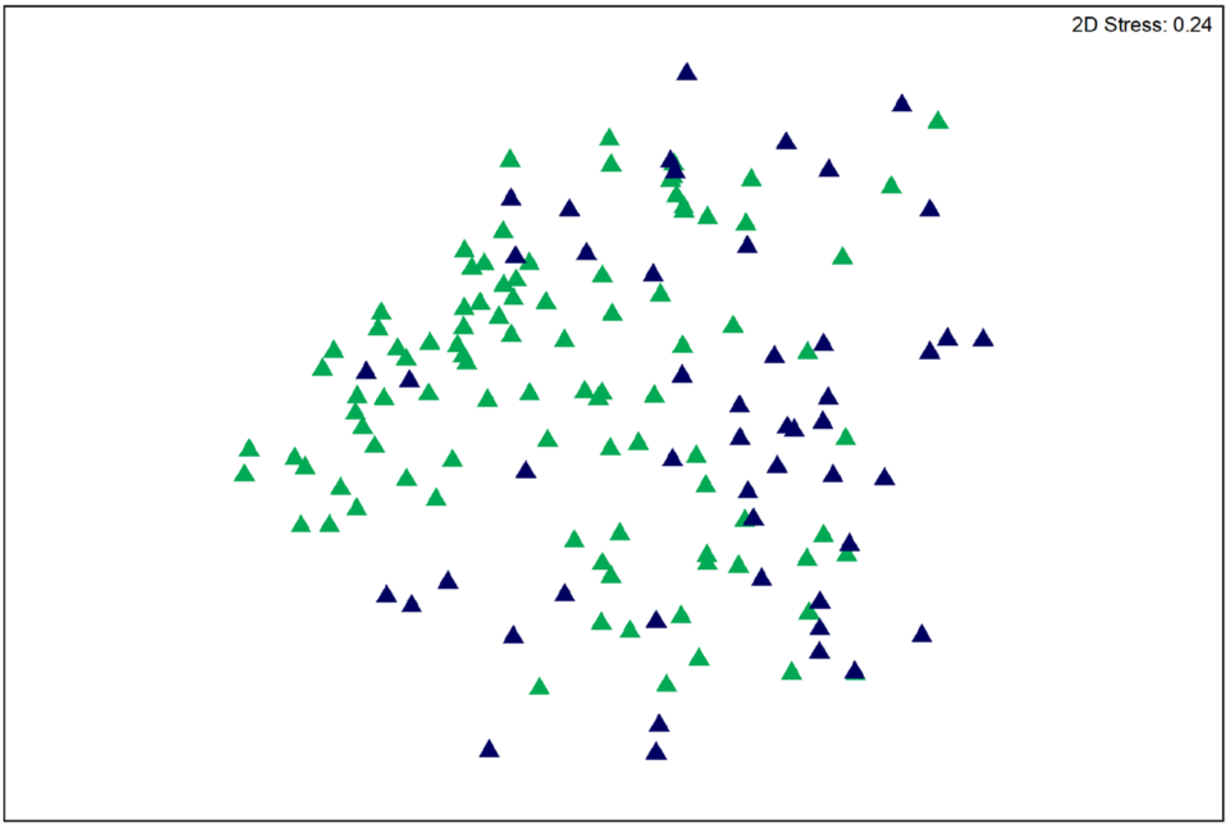


Figure 3.27 A two-dimensional MDS plot of the assessment sub-profile of educator perspectives on CT (n=146).Educators are coloured by the education setting they teach in, with the green triangles representing the high school educators and the dark blue triangles representing the tertiary educators.

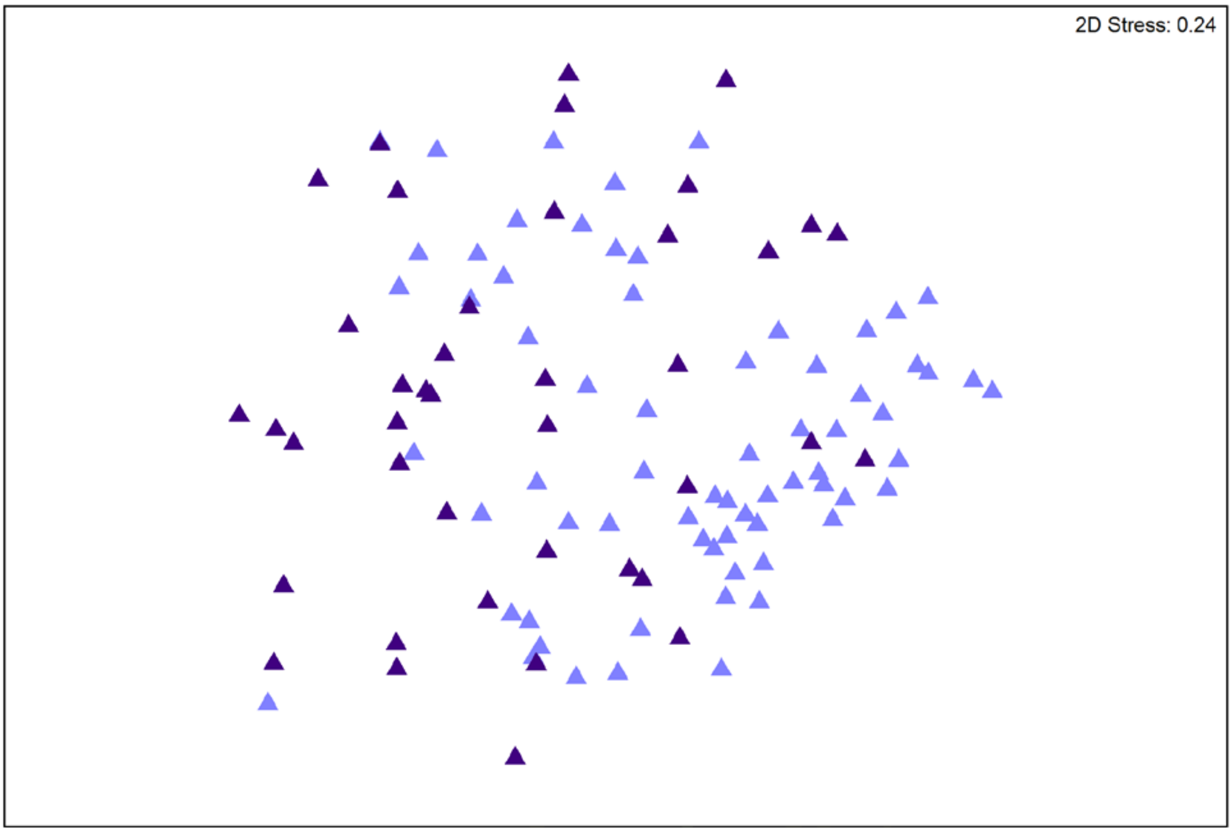


Figure 3.28 A two-dimensional MDS plot of the assessment sub-profile of science educator perspectives on CT (n=110). Educators are coloured by the education setting they teach in, with the light purple triangles representing the high school science educators and the dark purple triangles representing the tertiary science educators.

Table 3.15 Summary of the contributions of the items for each question, to factors with significant group differences in the assessment sub-profile.

respondent type	sig. level^		Question Set 5 (24 items)	Question Set 6 (9 items)
High school vs Tertiary	***	% contribution to group differences	88.32	11.66
		min - max %	2.2 - 4.59	1.87 - 4.59
Science: High school vs Tertiary	***	% contribution to group differences	87.95	12.06
		min - max %	0.49 - 3.53	0.58 - 3.56

^ significance level based on PERMANOVA results (refer to table 5.2);
*denotes p<0.05, **denotes p<0.01, ***denotes p<0.001

Table 3.16 Summary of the top 17 items contributing to group dissimilarity analysis from the SIMPER approach for demographic factors identified as significant in the assessment sub-profile. Displayed in rank order of contributions to the differences by education setting (i.e. between high school and tertiary educators). A complete list of the rank order of 33 sub-profile items dissimilarity contributions is shown in Appendix C: Table C.7.

Item theme	Item code	SIMPER rank (% contribution of item to difference between groups; group with higher mean score)			
		Education setting (HS vs TER)		Science: Education setting (Sc HS v Sc TER)	
self-developed course evaluation form	**A_5a - SD**	**1**	**(4.59%; HS)**	4	(4.49%; Sc HS)
self-developed discipline specific CT test	A_5b - SD	2	(4.57%; TER)	2	(4.58%; Sc TER)
externally sourced discipline specific CT test	A_5b - ER	3	(4.54%; HS)	3	(4.52%; Sc HS)
externally sourced general CT test	**A_5d - ER**	4	(4.54%; HS)	**1**	**(4.59%; Sc HS)**
self developed rubric	A_5g - SD	5	(4.43%; HS)	7	(4.32%; Sc HS)
would not use multiple ranking items	A_5e - NO	6	(4.36%; TER)	10	(4.06%; Sc TER)
self-developed self-report items	A_5h - SD	7	(4.33%; HS)	9	(4.23%; Sc HS)
self-developed extended response items	A_5c - SD	8	(4.21%; HS)	6	(4.34%; Sc HS)
would not use multiple choice items	A_5f - NO	9	(4.03%; TER)	15	(3.64%; Sc TER)
externally sourced extended response items	A_5c - ER	10	(3.98%; HS)	5	(4.42%; Sc HS)
would not use self-report items	A_5h - NO	11	(3.98%; TER)	12	(3.85%; Sc TER)
externally sourced multiple choice item	A_5f - ER	12	(3.93%; HS)	8	(4.23%; Sc HS)
would not use course evalaution form	A_5a - NO	13	(3.77%; TER)	14	(3.78%; Sc TER)
self-developed multiple choice items	A_5f - SD	14	(3.77%; TER)	13	(3.79%; Sc TER)
would not use general CT test	A_5d - NO	15	(3.76%; TER)	17	(3.41%; Sc TER)
externally sourced multiple-ranking items	A_5e - ER	16	(3.69%; HS)	11	(3.93%; Sc HS)
I look for specific evidence of CT embedded in students work.	A_6d	17	(3.53%; TER)	16	(3.56%; Sc TER)

3.4.4 Concerning Australian educator perspectives from the global profile: across all quantitative questions.

The main aim of this study was to explore the ideas represented in Australian educators' perceptions of CT. While standard survey analysis approaches would achieve this goal through comparing the findings across questions in the discussion, the multivariate approach used in this study meant that additional explorations across the quantitative items could be undertaken by generating a global profile. For this profile, educator perspectives about CT by the nature, development and assessment of CT across the 80 survey items were collated so perspectives could be considered holistically. Incidentally, no two respondent's global profiles were identical. After generating this global profile, three further analysis steps were taken, including hypothesis testing the global profile; exploring the extent of correlations across the sub-profiles and between the sub-profiles and global profile; and lastly, exploring the global profile for non-*a priori* groups. This section concludes with revisiting some findings with new non-*a priori* statistically-based groupings.

Results from the global profile

Differences in educator perspectives were explored by the main demographic factors (discipline, education setting, gender, state, teaching experience) through ordination and

PERMAONVA testing techniques, for all respondents (n=146) and in the science cohort (n=110). This was followed by dissimilarity analysis through the SIMPER process for factors identified as significant to help reveal which aspects of CT (aka which items) contributed to group differences.

When comparing by discipline, the MDS ordination revealed that there was variation in the perspectives, but history educators were more tightly clustered (Figure 3.29). PERMANOVA analysis revealed that the perspectives of science and history educators were indeed highly significantly different when considering their global profiles towards CT (Table 3.17: pseudo-F = 4.25, p-value = 0.0001). To investigate those items that contributed most to the discernible difference between the science and history educators' global profiles, each of the 80 items were ranked based on their percentage contribution to the overall difference between the discipline groups (using SIMPER analysis). In total, 14 of the 80 items contributed >2% to the difference between science and history global profiles. These 14 items belonged to three of the six question-sets (Table 3.18). The emerging perspectives concerned educators' development philosophies (Question 3), classroom approaches (Question 4) and views on CT assessment tools (Question 5). However, when considering the cumulative percentage contribution for all items within a question-set, only Question 4 and Question 5 contributed greatly (>70% of the cumulative scores) to the overall differences between global profiles of science and history educators (Table 3.19). This is in part due to both question-sets containing >18 items, as well as having a few items that contributed highly to this observed difference. For example, Question 5 cumulatively had 10 out of 24 items that contributed highly (>2%) to the global profiles. Even so, the global profile was constructed such that all items had equal weighting, and this meant questions with only a few items could still contribute highly to group differences. For example, two of the three items from Question 3 (items D_3a and D_3b) contributed as much to differences in the global profile as the highest contributing items from Question 4 and 5 (Table 3.18), and thus Question 3 also had some weight at discerning the overall differences between the perspectives held by the two disciplines. However, it was perspectives about the scientific method (theme of item D_4h) that contributed the most to group differences in the global profile when exploring by the demographic factor discipline (Table 3.18). In addition, item D_4i (Socratic questioning) also stood out among those items which contributed at least 2% to group differences because it was only important for discerning differences by discipline (and not other demographic factors). For this item, fewer science than history educators indicated they would use this development approach.

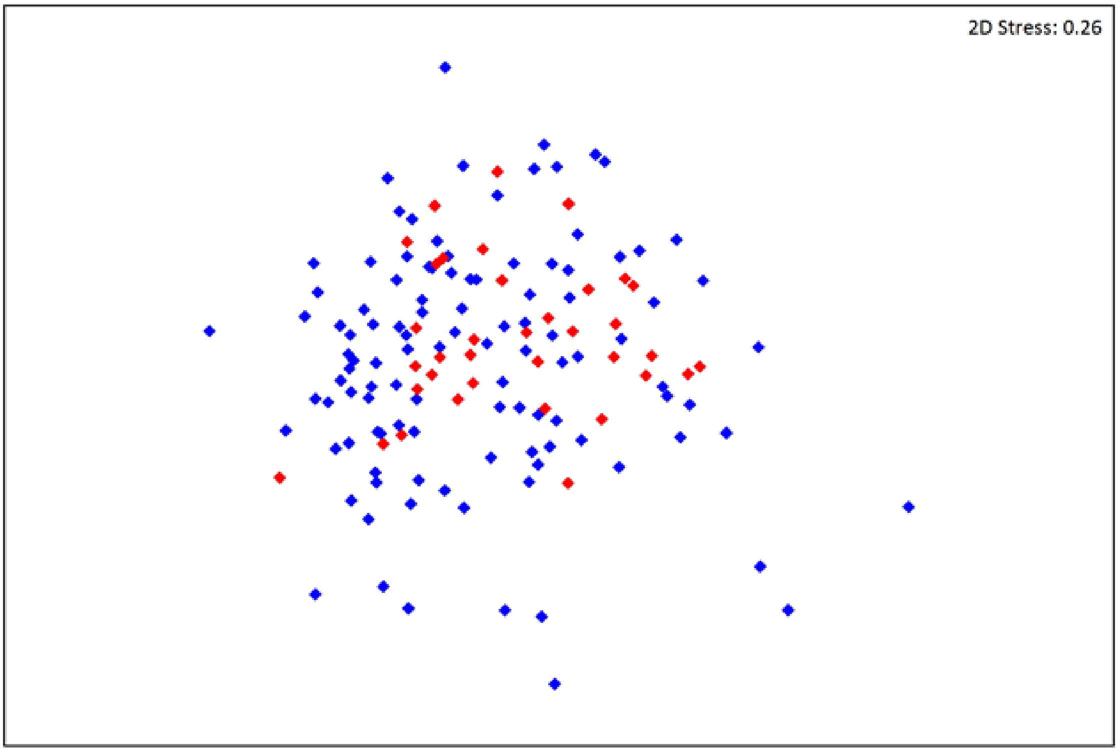


Figure 3.29 A two-dimensional MDS plot of the global profile of educator perspectives on CT. Individuals are coloured by the discipline they teach, with the red diamonds representing the history educators and blue diamond's representing the science educators.

Table 3.17 Summary of findings for *a priori* hypothesis tests undertaken on the global profile, using PERMANOVA to explore if any particular demographic factors underpinned educator perceptions about CT.

Factor	global profile		
	pseudo-F	p-value	df
discipline	**4.25**	**0.0001**	**1**
education setting	**7.45**	**0.0001**	**1**
gender	1.77	0.0645	1
state	1.11	0.2677	6
teaching experience	1.39	0.1193	2
discipline vs education setting	0.68	0.7444	1
discipline vs gender	1.12	0.3416	1
discipline vs state*	1.08	0.3345	4
discipline vs teaching experience	0.87	0.6165	2
education setting vs gender	0.52	0.8782	1
education setting vs state*	1.63	0.0106	4
education setting vs teaching experience	0.72	0.8054	2
gender vs state*	0.95	0.5717	6
gender vs teaching experience	1.07	0.3768	2
state* vs teaching experience	0.95	0.6044	8
science - education setting	**6.41**	**0.0001**	**1**
science - gender	1.49	0.1326	1
science - state	1.19	0.1554	6
science - teaching experience	1.14	0.3031	2

*pairwise tests incorporating state as a factor excluded NT and ACT respondants from analysis due to low sample size.

Table 3.18 Items which contributed >2% to global differences (as per SIMPER test) between respondents by discipline, education setting and by education setting within the science only cohort. Bolded codes, themes and x's indicate the items that contributed the most to differences by that demographic comparison. The capital letter in the item code refers to the sub-profile to which the item belongs. For example, D = development sub-profile, and A = assessment sub-profile. A full list of item codes and corresponding themes/statements is found in Appendix B.

		Demographic factor		
Item code	Item theme	Discipline (n = 146)	Education setting (n = 146)	Science: Education setting (n = 110)
D_3a	explicit instruction	x	x	
D_3b	embedded instruction	x	x	x
D_4c	concept mapping			x
D_4h	**scientific method**	**x**		
D_4i	Socratic questioning	x		
A_5a - SD	**self-developed course evalution form**	x	**x**	x
A_5b - ER	use external resource: discipline specific CT test	x	x	x
A_5b - SD	self-developed discipline specific CT test	x	x	x
A_5c - ER	use external resource: extended response items		x	x
A_5c - SD	self-developed extended response items	x	x	x
A_5d - ER	use external resource: general CT test	x	x	x
A_5e - ER	**use external resource: multiple-ranking items**			**x**
A_5e - NO	would not use multiple-ranking items	x	x	x
A_5f - ER	use external resource: multiple choice items		x	x
A_5f - NO	would not use multiple choice items	x	x	
A_5g - SD	self-developed rubric	x	x	x
A_5h - NO	would not use self-report items	x	x	
A_5h - SD	would self develop self-report items	x	x	x

Table 3.19 Summary of the group differences for demographic factors identified as significant in the global profile. Percentage contribution based on the dissimilarity values from the SIMPER test. The cumulative score for each question set is calculated from the sum of individual items. The min-max % contribution indicates the range of possible contributions per item from that set.

			Question Set#					
respondent type	sig. level^		1 (11 items)	2 (15 items)	3 (3 items)	4 (18 items)	5 (24 items)	6 (9 items)
Science vs History	***	% contribution to group differences	4.85	10.55	4.95	29.14	44.62	5.87
		min - max %	0.06 - 0.9	0.41 - 0.94	0.43 - 2.27	0.98 - 2.58	0.66 - 2.31	0.24 - 1.76
High school vs Tertiary	***	% contribution to group differences	4.75	10.46	4.87	29.04	44.93	5.92
		min - max %	0.06 - 0.87	0.45 - 0.9	0.64 - 2.17	1.07 - 1.93	1.17 - 2.31	0.24 - 1.77
Science: High school vs Tertiary	***	% contribution to group differences	4.81	10.57	4.88	28.71	44.95	6.13
		min - max %	0.07 - 0.89	0.45 - 0.91	0.78 - 2.14	1.11 - 2.02	0.98 - 2.31	0.28 - 1.79

number of items in question set displayed in brackets below question number

^ significance level based on PERMANOVA results (refer to Table 4.17); *denotes p<0.05, **denotes p<0.01, ***denotes p<0.001

When exploring for differences in perspectives by education setting, the global profiles were examined for differences across all 146 respondents, and within the 110-science respondent cohort. Both ordinations revealed there was slightly less diversity in the high school educator perspectives (Figure 3.30 and Figure 3.31), where these educators were more tightly clustered than the tertiary educators were (and therefore showed more similarity in their perspectives). When investigating education setting differences using PERMANOVA, the global profiles of high school and tertiary educators were revealed to be highly significantly different (Table 3.17: pseudo-F = 7.45, p-value = 0.0001), as were the global profiles of high school science and tertiary science educators (Table 3.17: pseudo-F = 6.41, p-value = 0.0001).

For the 146-respondent cohort, the SIMPER test revealed that 14 out of the 80 items contributed >2% to the discernible differences between the global profiles of high school and tertiary educators (Table 3.18). These items belonged to Question 3 (development perspectives) and Question 5 (assessment tools). However, when considering the cumulative percentage contribution of each question-set to the global profile, only Question 5 contributed greatly (>44% of the cumulative scores) to the observed differences between high school and tertiary respondents (Table 3.19). Item 5a – SD (course evaluation form) contributed the most to group differences by education setting, where more high school educators indicated they would use a self-developed version of this tool type than tertiary educators. This item was also important for group differences in by discipline and within the science education cohort. In fact, 12 of the 14 items also contributed to the observed differences between science and history respondents (refer back to Table 3.18). Of these items that contributed at least 2% to group differences, perspectives on externally sourced assessment tools contributed more to differences by education setting (including within the science cohort) than by discipline. For example, items A_5c-ER (extended response items) and A_5f-ER (multiple-choice items) contributed highly to group differences by education setting (both cohorts) but not by discipline (Table 3.18). In both instances, high school educators were more likely to seek out and use this assessment resource type.

Within the science cohort, thirteen discriminating items contributed greater than 2% to group differences (Table 3.19). These came from Question 3 (development perspectives), Question 4 (classroom approaches) and Question 5 (assessment tools). Nine of these items also contributed highly to overall differences by education setting and discipline (Table 3.18). However, there were two items (D_4c, A_5e_ER) which contributed more to education setting group differences in the science cohort (n=110), than in the overall education setting or discipline analyses (n=146). Item 5e –ER (externally sourced multiple-ranking items) contributed the most to group differences within the science cohort, where more science high school educators indicated they would seek out and use this assessment tool type.

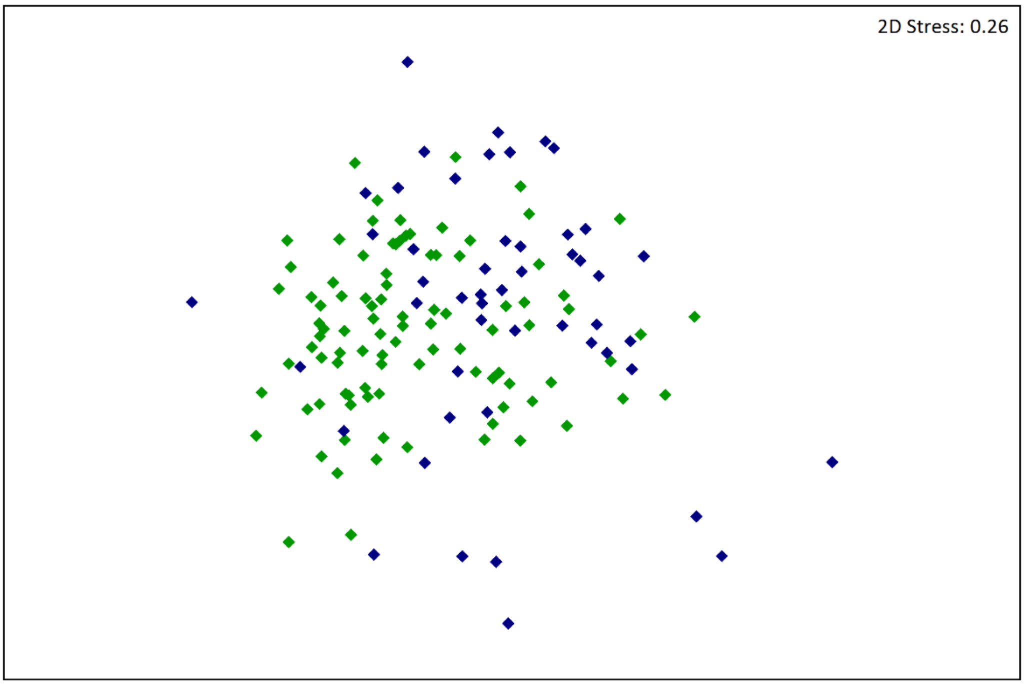


Figure 3.30 A two-dimensional MDS plot of the global profile of educator perspectives on CT (n=146). This is the same MDS plot as Figure 3.29, however, in this figure educators are coloured by the education setting they teach in, with the green diamonds representing the high school educators and the dark blue diamond's representing the tertiary educators.

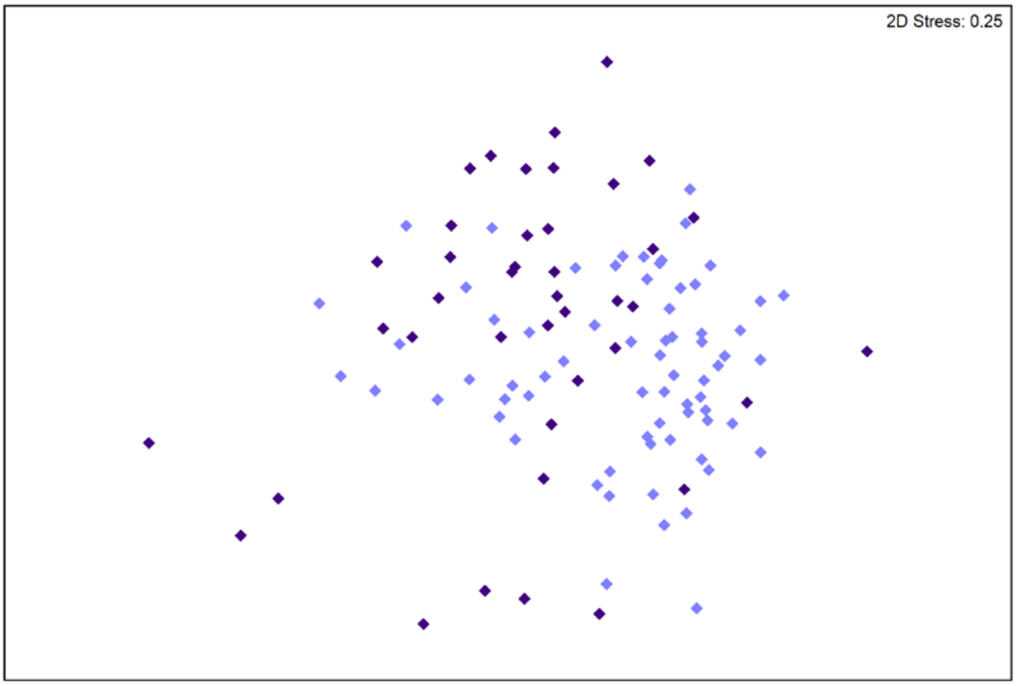


Figure 3.31 A two-dimensional MDS plot of the global profile of science educator perspectives on CT (n=110). Science educators are coloured by the education setting they teach in, with the light purple diamonds representing the high school educators and the dark purple diamonds representing the tertiary educators.

There were no significant differences observed at the 0.001 level for gender, state, or years of teaching, in the global profile (irrespective of the presence of the history outgroup respondents; Table 3.17). However, since this is an exploratory study, there were a few other results in the global profile that sat on the cusp of moderate to high significance that warrant brief consideration. For example, although there were no observed differences between the states themselves, there was an interaction effect between states versus education setting (Table 3.17: pseudo-F = 1.63, p-value = 0.0106). This implied that the pattern of perspectives within one state is not reflected in other states. The pairwise comparisons showed that there were only highly significant differences between high school and tertiary educators within the states of VIC; and high school and tertiary educators within NSW (see footnotes Appendix C: Table C.2). There were also significant differences in the perspectives held by high school educators from VIC and SA; and between the perspectives of high school educators from SA and WA (see footnotes Appendix C: Table C.2). SIMPER analysis revealed that Question 5 had the highest contributing item for three of the four interaction effect pairings (see Question 5 max percentage contributions in Table 3.20). However, the item contributing the most to the group differences varied depending on whether the comparison was across states, or within a state. For example, perspectives on self-developing a rubric to assess CT (theme of item A_5g-SD) contributed the most to group differences when comparing between high school educator perspectives in SA and WA. Specifically, more WA high school educators indicated they would self-develop this assessment tool than SA high school educators. This same item also contributed the most to group differences between high school educators from SA and VIC; however, in this instance more SA educators than VIC high school educators indicated they would self-develop this kind of tool to assess CT. There were also five other highly contributing items which were unique to differences between states at the high school level (Appendix C Table C.8: D_3a, D_4i, A_5d-NO, A_5e-NO, A_5f-ER). There were also three items which only contributed highly to the differences between SA and WA high school educators (Appendix C Table C.8: items A_5d-NO, A_5e-SD, A_5f-SD).

Perspectives on the use of self-developed course evaluation forms (theme of item A_5a-SD) contributed the most to group differences between high school and tertiary educators from NSW. Perspectives on A_5a-NO were also important for defining differences in this interaction effect, but not for other interaction effect group differences (Appendix C Table C.8). In this instance, more tertiary educators than high school educators indicated they would not use course evaluation forms to assess CT. For the remaining interaction effect, an item from Question 4

(D_4r: argument analysis) contributed the most to group differences between Victorian high school and tertiary educators. In addition, item D_4r and three other items only contributed highly to differences between Victorian educator perspectives, and not (highly) to other interaction effect pairings (Appendix C Table C.8: items 4k, 4l, 4r and 5c-NO).

Overall, these findings indicate that aspects of CT development and assessment play more important roles in discerning differences between educator ideas about CT than conceptual elements. While there were highly significant differences between some educator groups in the global profile (i.e. discipline and education setting), there was also variation in the way educators responded, and therefore no total separation between groups by these demographic factors in their ordinations. This indicated that demographic factors do not solely drive choices relating to CT.

Table 3.20 Summary of the group differences for state versus education setting; these interaction effect pairings were identified as significant in the global profile. Percentage contribution based on the dissimilarity values from the SIMPER test. The cumulative score for each question set is calculated from the sum of individual items. The min-max % contribution indicates the range of possible contributions per item from that set.

respondent type	sig. level^		Question Set# 1 (11 items)	2 (15 items)	3 (3 items)	4 (18 items)	5 (24 items)	6 (9 items)
Victoria: High school vs Tertiary	***	% contribution to group differences	4.39	10.14	4.01	30.37	45.23	5.85
		min - max %	0 - 1.06	0.16 - 1.07	0.2 - 1.94	0.98 - 2.92	0.79 - 2.63	0.04 - 2.1
NSW: High school vs Tertiary	***	% contribution to group differences	4.8	9.96	5.11	29.91	43.81	6.41
		min - max %	0.05 - 0.98	0.42 - 0.91	1.09 - 2.13	1.34 - 1.93	0.68 - 2.56	0.15 - 1.49
High school: SA vs Vic	**	% contribution to group differences	4.69	10.42	5.73	23.35	49.42	6.45
		min - max %	0 - 1.03	0.3 - 0.99	0.61 - 2.65	0.64 - 2.03	0.45 - 3.17	0.12 - 2.22
High school: SA vs WA	*	% contribution to group differences	4.97	10.72	5.3	23.16	49.88	5.96
		min - max %	0 - 0.86	0.24 - 1.05	0.41 - 2.55	0.6 - 2.14	0.42 - 3.3	0 - 1.93

number of items in question set displayed in brackets below question number

^ significance level based on PERMANOVA results (refer to Table 4.17); *denotes p<0.05, **denotes p<0.01, ***denotes p<0.001

Comparisons of the contributions of the sub-profiles to the global profile using Mantel's test

Exploration across the educator responses has highlighted that there were differences within the global and sub-profiles. Is there any connection between these differences? That is to ask if a particular result was observed in an aspect of the development profile, did this result correlate to a particular conception of CT or trend in the assessment of CT? Since the global profile comprises of three theoretical elements related to CT, it was possible to assess whether a pattern derived from one of the sub-profiles of CT (such as developmental aspects) is reflected in the other sub-profiles of CT (i.e. assessment and/or conceptual aspects). To undertake this analysis, each of the patterns produced from the three sub-profiles (generated from 146 respondents) were correlated (using the Mantel's test). There was a mild, yet significant correlation between the development and assessment sub-profiles, and each contributed highly and equally to the global profile (with rho values >0.7, Figure 3.32). However, the contribution of the conceptual sub-profile to the global profile was less (rho=0.117), showing only a mild correlation. Even though the conceptual profile was comprised of fewer items (11/80) than the other sub-profiles (Figure 3.32: 33 items and 36 items), its overall contribution to the global profile (rho=0.117) was less than could be solely explained by the smaller number of items. Furthermore, there was little if no correlation between the conceptual sub-profile and the assessment and development sub-profiles. This indicates that while respondents may hold a particular perception of CT reflected in their assessment and development approaches, they may have differing conceptual ideas.

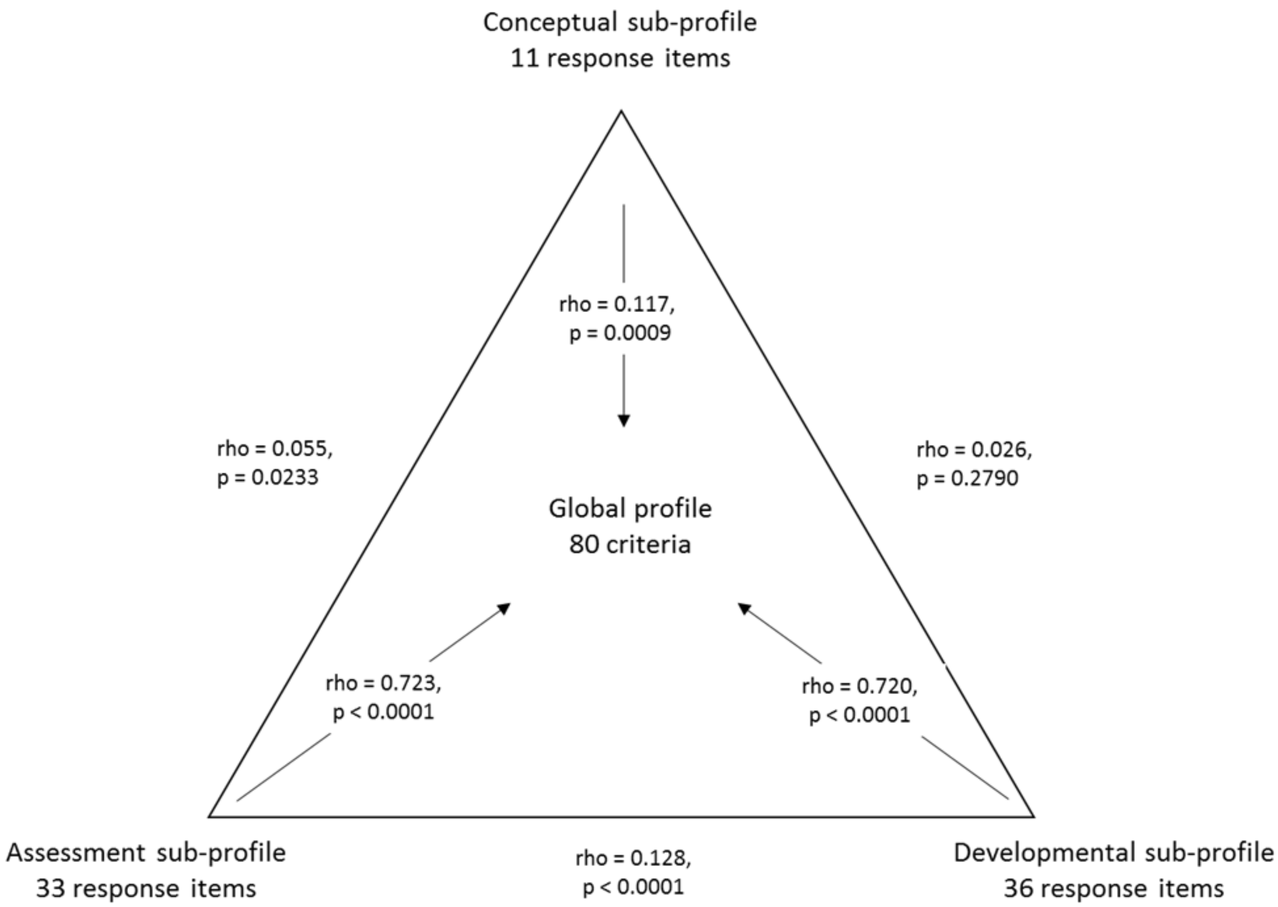


Figure 3.32 Summary of the Spearman correlation coefficients of a partial Mantel test results between global and sub-profiles for educator responses.

Given that the strongest correlation was between assessment and development, the items responsible for this correlation were explored using vector overlays on the assessment sub-profile

PCoA ordination. This showed that when educators responded with particular approaches to assessment items (Figure 3.33), some vectors from the sub-development profile also pointed in the same direction (Figure 3.34). This indicates a similarity in the way those educators view aspects of CT development and assessment. For example, educators that indicated they were likely to self-develop assessment tools (i.e. see respondents near the top of the data cloud in Figure 3.33 where the vectors for items A_5b-SD, A_5d-SD, A_5f-SD, A_5h-SD are pointing) also shared a perspective about how effective problem-based learning (D_4a) for developing CT (Figure 3.34).

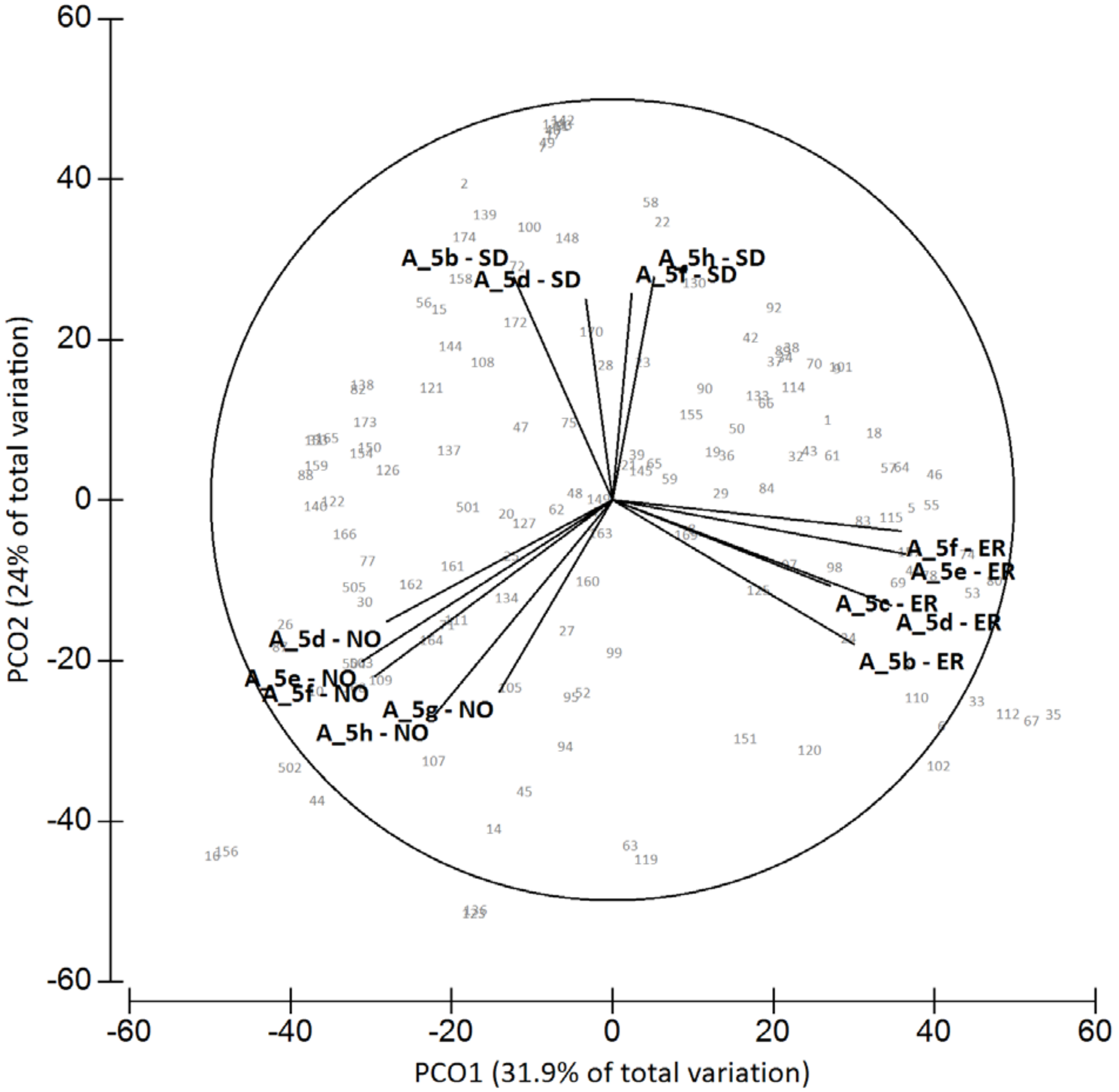


Figure 3.33 Two dimensional PCoA of educator responses to the assessment sub-profile with assessment profile vector overlay (n=146).The length of each vector line indicates the extent of correlation of the responses to items (minimum correlation displayed 0.5), in addition to collectively highlighting what respondents positioned in that part of the data cloud answered similarly. The capital letter in the item code refers to the sub-profile to which the item belongs. For example, D = development sub-profile, and A = assessment sub-profile. A full list of item codes and corresponding themes/statements is found in Appendix B.

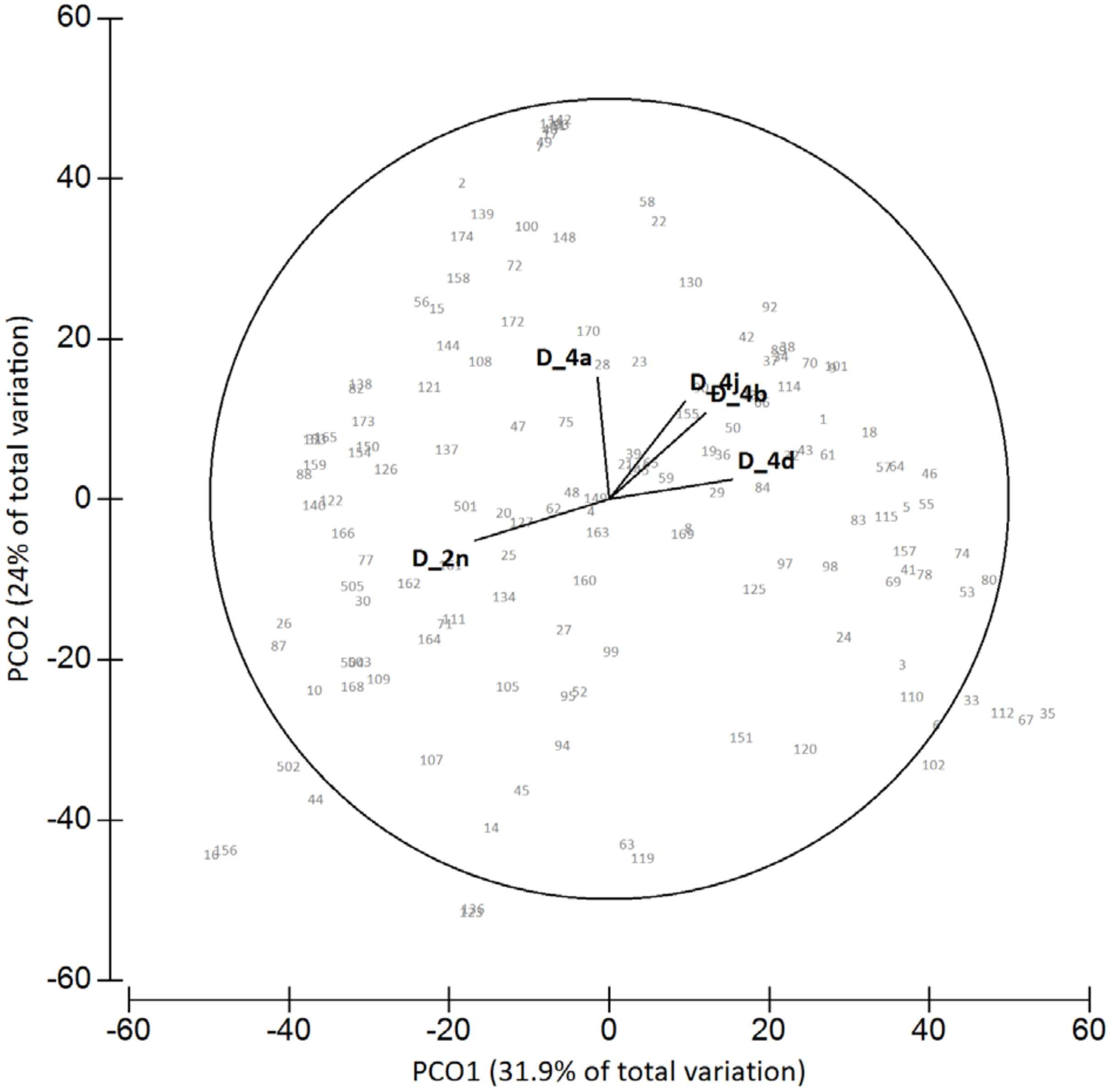


Figure 3.34 Two dimensional PCoA of educator responses to the assessment sub-profile with vectors of development sub-profile items overlayed on the assessment ordination (n=46). The length of each vector line indicates the extent of correlation of the responses to items (minimum correlation displayed 0.3), in addition to collectively highlighting what respondents positioned in that part of the data cloud answered similarly. Reminder D= items from the development sub-profile. A full list of item codes and corresponding themes/statements is found in Appendix B.

Exploring for* non-*a priori groups and describing them.

Although no two respondents had the same global profile, there were statistically derived clusters of respondents (using the SIMPROF algorithm at a 5% significance level). The 146 respondents formed 23 groups, 11 of which comprised of greater than four respondents. These 11 groups accounted for 80% of the respondents (Table 3.21). While some of the groups could be attributed to being high school educators (i.e. Group G and J), some groups (like Groups N, U, V), contained a mix of educators from different disciplines, states, gender, education settings and years of teaching experience (Table 3.21). This highlights that while key educator background characteristics could explain some of the variability in profiling perception of CT, there are additional and unaccountable attributes that define the reasons why educators hold a certain perceptive.

When exploring the items that contributed highly to the global profile using the *a priori* groups, items from three questions (Questions 3, 4 and 5) regularly came up as contributing highly to group differences. However, when using the non-*a priori* groups, five questions were important for defining group characteristics (Questions 2 and 6 in addition to Questions 3, 4, and 5 see Table 3.21). This increase in the number questions contribution to group difference these meant the non-*a priori* groups helped clarify the sources of variation in Australian educator perspectives about CT. For example, item D_2g (intercultural understanding) was the item that respondent's in Group A answered most similarly, and therefore helped set apart these educator's perspectives from the perspectives of educators in other groups (Table 3.21). Items from Question 5 were revealed to be the highest contributing item to group similarity for six of the eleven groups. In fact, every group had at least one item from Question 5 in their top five contributing items to group similarity, and the top five similarities for Group N were all from Question 5. The role of Question 5 in contributing to the general shape of the ordination (Figure 3.35) is particularly evident in Figure 3.36. The direction of the vectors in this figure highlight that preference for type of tool and the source of the assessment tool (general theme of Question 5 items) were responsible for separating the groups. For example, Group G (orange), P (maroon) and N (yellow) form three of the five boundaries in the data cloud; Group G – were similar in their responses to self-developed assessment styles, Group P – were similar in their response to not using particular assessment tools, and Group N – similar in their responses to seeking out and using external assessment tools. In addition, the longest vectors correspond to items from Question 5 (Figure 3.36).

Table 3.21 Summary of key group traits and their most shared CT perspectives for SIMPROF groupings. The sub-profile the item belongs to is signified by the capital letter in the item code. For example, D = development sub-profile, and A = assessment sub-profile. A full list of survey questions and items can be found in Appendix B.

Group ID	Group size	Discipline		Education setting		Perception of development strategy			Common* characteristics for this group by item codes+				
	n	Science (%)	History (%)	High school (%)	Tertiary (%)	explicit (%)	embedded (%)	no plan (%)					
J	28	82	18	96	4	25	75	0	D_4a	A_5g^	A_5d#	A_5a^	D_4l
V	20	70	30	40	60	10	90	0	D_3b	D_4a	A_5c^	D_2b	D_4n
N	11	64	36	82	18	27	55	18	A_5a^	A_5c^	A_5d^	A_5h^	A_5g^
O	10	60	40	70	30	90	10	0	A_5c^	D_3a	D_2n	D_4f	D_2f
U	10	50	50	50	50	10	90	0	A_5b~	A_5e~	A_5f~	D_3b	A_5d~
G	9	78	22	100	0	33	67	0	A_5a#	A_5d#	A_5e#	A_5h#	D_2b
I	8	88	13	75	25	75	13	13	A_5b#	A_5c^	A_5d#	A_5g^	D_4n
A	7	100	0	14	86	71	0	29	D_2g	A_5h~	A_5b#	A_5d#	D_2n
C	6	100	0	0	100	17	83	0	A_5b^	D_4a	D_2b	D_2n	A_6d
P	5	100	0	80	20	100	0	0	D_3a	A_5a~	A_5c^	A_5e~	A_5g^
T	5	20	80	60	40	100	0	0	D_3a	A_5e~	A_5f~	A_5h~	D_4r
other	27	89	11	63	37	30	52	19	mixed ideas due to n<5 in remaining groups				

*Common refers to the most consistent highest scoring items among members of that group as determined by simper analysis; descriptions of items provided in text in Table 4.22.

+ To make it easier to compare across the items contributing the most to group identity, the item codes have been shortened. For example item A_5b (course evaluation forms) was the highest contributing item to group identity of three groups (n, u and c,) however each of these groups took a different approach in how they would engage with this assessment tool.

^Self-developed (-SD)

#External resource (-ER)

~Would not use (-NO)

Table 3.22 Descriptions of items relating to the most shared CT perspectives for SIMPROF groupings. A full list of survey questions and items can be found in Appendix B.

Group ID	Group size *n*	Common* characteristics for this group
J	28	report problem based learning as effective approach; would use a self-developed rubric to assess CT; would use externally developed general CT test to assess CT; would use a self-developed course evaluation form to assess CT; report deductive reasoning as an effective technique they have used for for developing CT
V	20	employs embedded development approaches; report problem based learning as an effective technique they have used for for developing CT; would use a self-developed extended response items to assess CT; viewed CT as much more important than citizenship; report class discussions as effective for developing CT
N	11	would use self-developed course evaluation forms to assess CT; would self-develop extended response items to assess CT; would self-develop a general CT test to assess CT; would use a self-developed self-report items to assess CT; would self-develop a rubric to assess CT
O	10	would use self- developed extended response items to assess CT; explict development approaches; viewed CT as much more important than teamwork; report debating as an effective technique they have used for for developing CT; viewed CT as much more important than ICT
U	10	would not use self-developed CT tests to assess CT; would not use multiple-ranking items to assess CT; would not use multiple-choice items to assess CT; employs embedded development approaches; would not use a general CT test to assess CT
G	9	would use an externally developed course evaluation form to assess CT; would use an externally developed general CT test to assess CT; would use externally developed multiple-ranking items to assess CT; would use an externally developed self-report items to assess CT; viewed CT as much more important than citizenship
I	8	would use an externally developed discipline-specific CT test to assess CT; would self-develop extended response items to assess CT; would use an externally developed general CT test to assess CT; would self-develop a rubric to assess CT; report class discussions as effective for developing CT
A	7	viewed CT as much more important than intercultural understanding; would not use self-report items to assess CT; would use an externally developed discipline-specific CT test to assess CT; would use an externally developed general CT test to assess CT; viewed CT as much more important than teamwork
C	6	would self-develop a discipline-specific CT test to assess CT; report problem based learning as an effective technique they have used for for developing CT; viewed CT as much more important than citizenship; viewed CT as much more important than teamwork; look for specific-evidence of CT embedded in their students work.
P	5	employs explicit development approaches; would not use a course evaluation form to assess CT; would self-develop extended response items to assess CT; would not use multiple-ranking items to assess CT; would self-develop a rubric to assess CT
T	5	employs explicit development approaches; would not use multiple-ranking items to assess CT; would not use multiple-choice items to assess CT; would not use self-report items to assess CT; report argument analysis as an effective technique they have used for for developing CT
other	27	all unique

*Common refers to the most consistent highest scoring items among members of that group as determined by simper analysis.

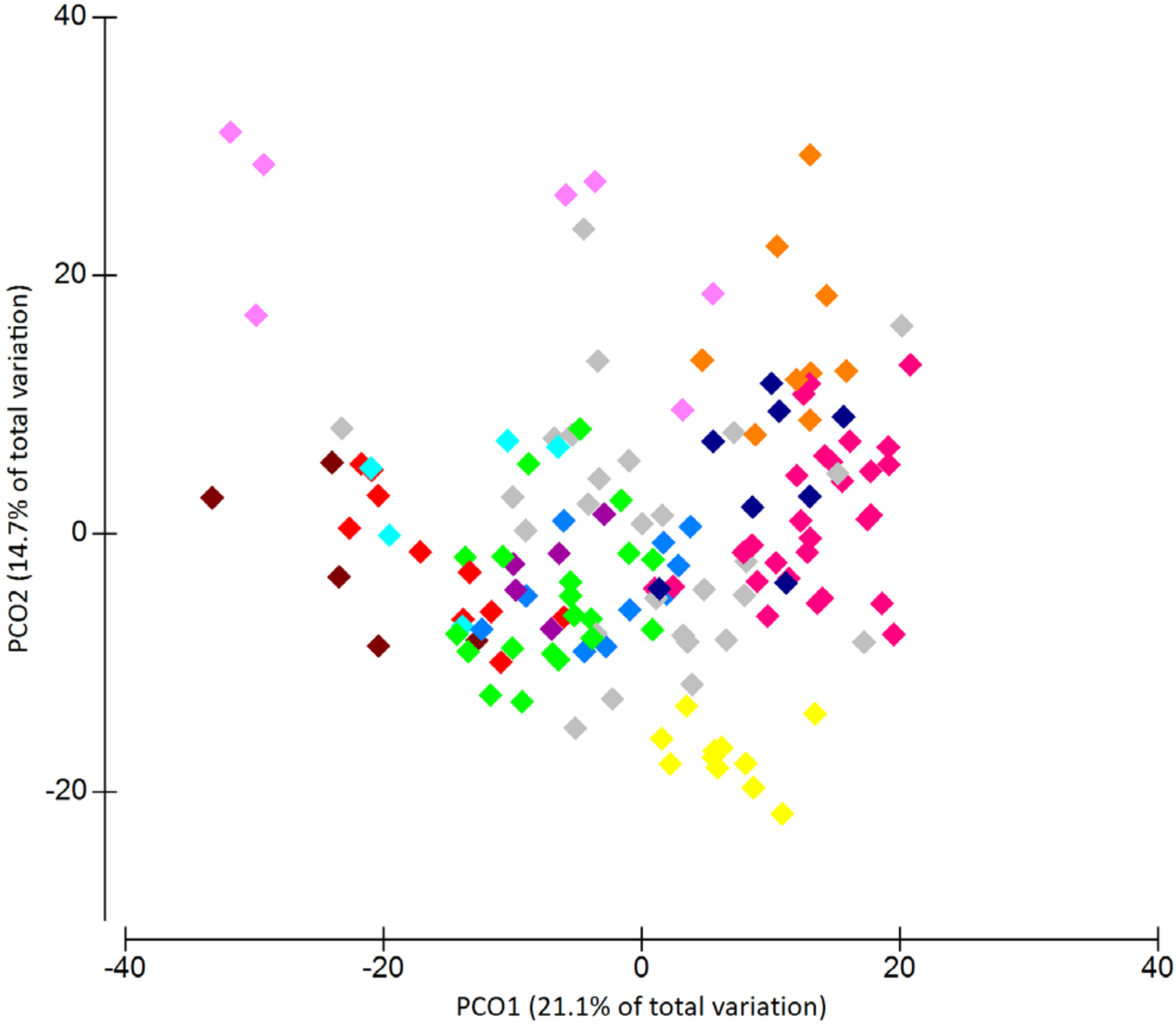


Figure 3.35 Two dimensional PCoA of educator responses to the global profile coloured by SIMPROF groupings (refer to Table 3.21 for group labels for the colours).

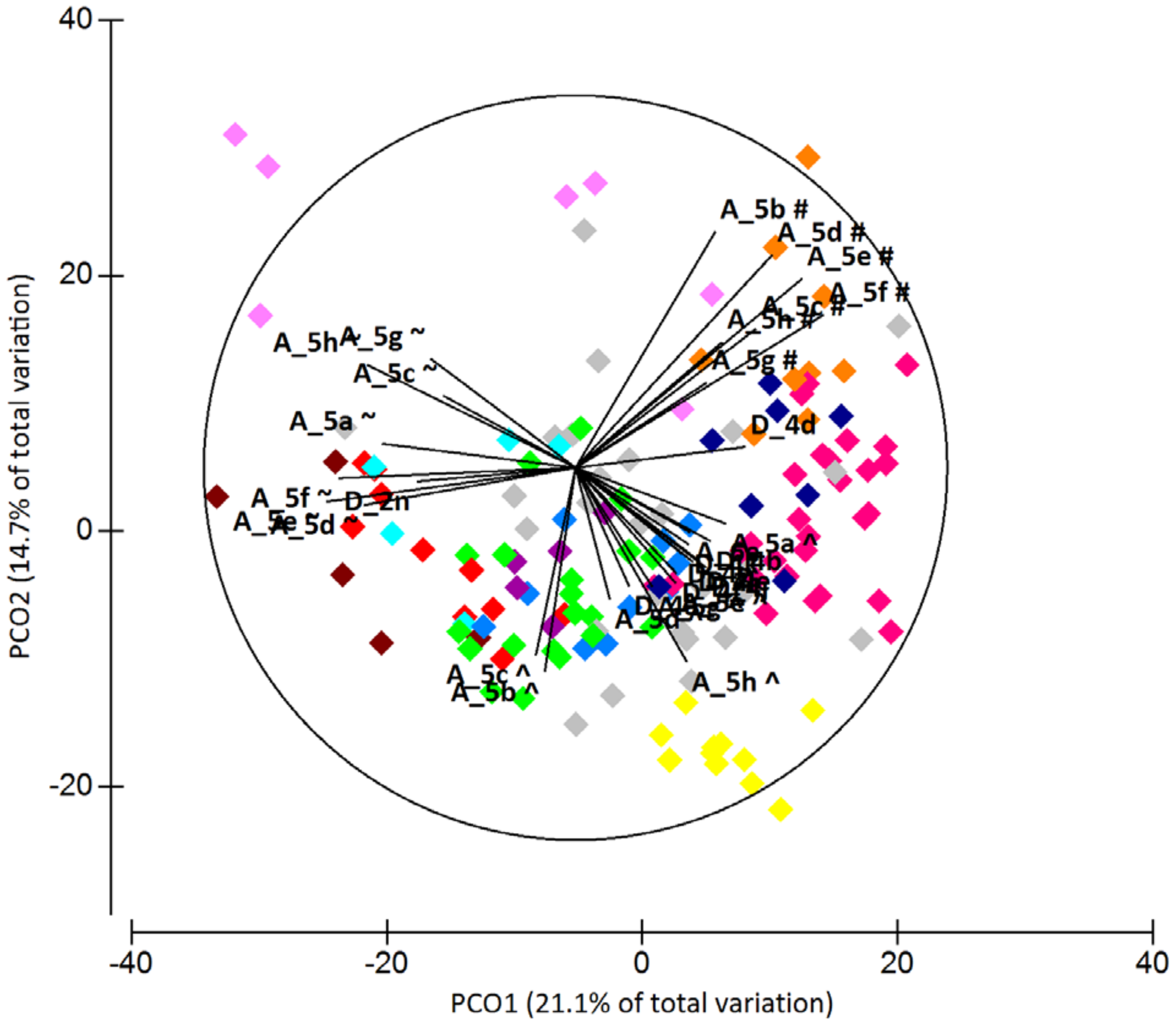


Figure 3.36 Two dimensional PCoA of educator responses to the global profile coloured by SIMPROF groupings with vectors from the global profile overlayed on the global ordination. The length of each vector line indicates the extent of correlation of the responses to items (minimum correlation displayed 0.35), in addition to collectively highlighting what respondents positioned in that part of the data cloud answered similarly. The capital letter in the item code refers to the sub-profile to which the item belongs. For example, D = development sub-profile, and A = assessment sub-profile. As per Table 3.21, the symbols represent themes relating to each assessment tool, i.e. ~Would not use ^Self-developed #external resource. A full list of item codes and corresponding themes/statements is found in Appendix B.

<u>Contribution of non-*a priori* groups to conceptual understanding</u>

Exploration of the global profile and sub-profiles using the *a priori* demographic factors revealed that the conceptual profile contributed little to the overall understanding about educator differences. Since the non-*a priori* groups were able to increase understanding about educator differences at a global level, was it also able to do the same a conceptual level?

A MDS ordination was used to exploring the trends among educator point spending and statement choices to Question 1 using the SIMPROF groupings (Figure 3.37). The ordination has a large central cluster that is made up of respondents from a mix of all the SIMPROF groups, and it is evident the respondents on the fringe of the data cloud come from a range of different groups.

However, there are more respondents from the yellow group on the fringe of the cloud than from any other grouping. Further, the respondents from Group U (red) and Group V (light green) appear to be more tightly clustered than any of the other groupings (except for one outlier in each). In addition, Group V also seems to split into two smaller clusters (with one outlier), where one of the clusters spent their points on items similar to the majority of other respondents, and the other cluster spent it on statements that were chosen less frequently (yet still similar to other members of Group V). Thus, the SIMPROF grouping has added a little more clarity to the understanding of conceptual differences, even though no exclusive clusters emerge.

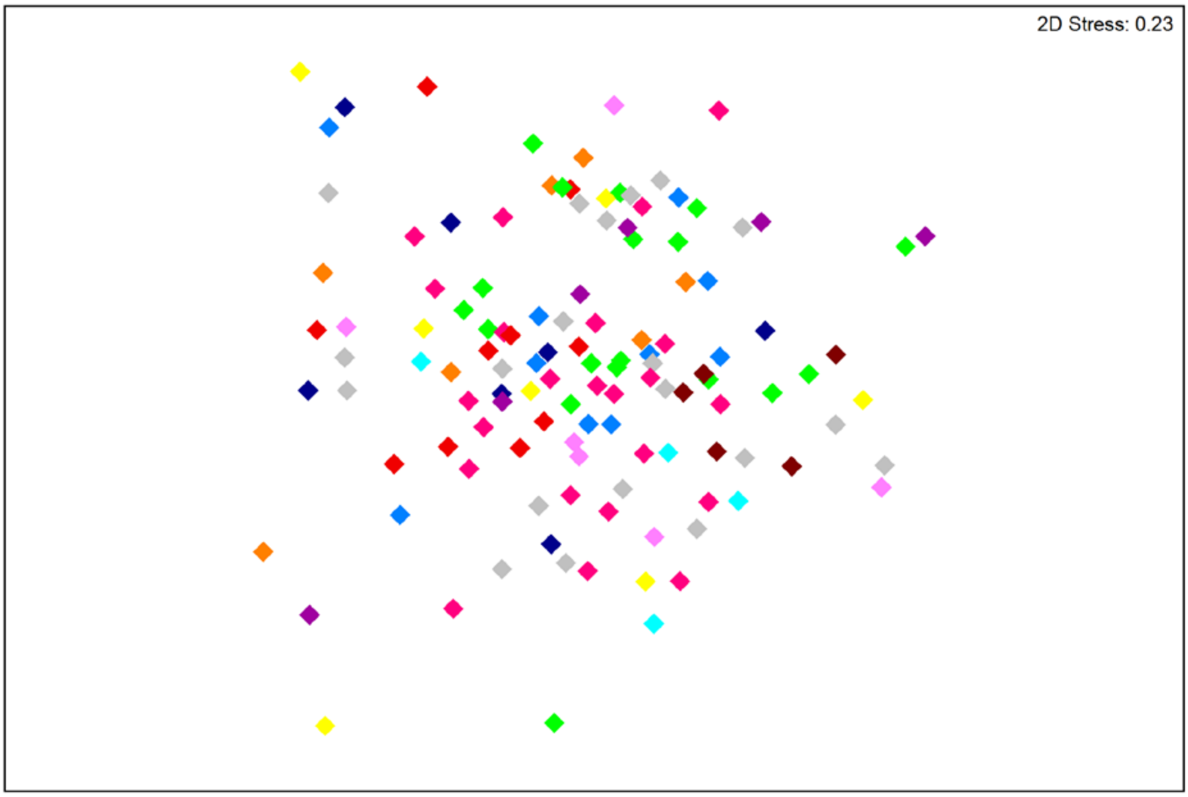


Figure 3.37 Two-dimensional MDS plot of the conceptual sub-profile of educator perspectives on CT (n=146). Educators are coloured by the grouping they clustered in from the SIMPROF analysis (refer to Table 3.21 for group labels for the colours).

When exploring the responses to Question 1a (qualitative) with the SIMPROF groupings (which are based on statistically significant similarities in the way that educators answered the quantitative items in the global profile), some clearer trends emerge than when exploring responses with the statistically significantly different *a priori* groups (see Table 3.23). For example, examining the inclusion of the framework element 'using information processing skills,' grouping by *a priori* factors portrays the perspective that the majority of educators (>85%) included this theme in their response, irrespective of the respondent's discipline or education setting. Yet examining the qualitative responses with the SIMPROF groups actually helped to identify those educators who did not include this theme in their response (i.e. Group G), and also provides insight into other shared differences in that group's perspective, compared to other educators in the study. Similarly, when exploring the inclusion of 'dispositions' in educator responses grouping by the significant *a priori* factors portrays the perspective that there is little difference between the general way that these groupings of educators included this theme in their response. Whereas the SIMPROF groupings highlight that there are particular subsets of educators (i.e. Group T and Group G) were more likely to include this theme in their response. What were the other key trends the SIMPROF groupings revealed?

Purposeful querying was a theme present in every groups' response, with Group T including it the most (67%), and Group P including it the least (25%). Aspects of the critical reflection element were mentioned in all but Group P's responses, and again they differed the most to Group T, where 100% of educators in this group included something related to this element in their response. Themes related to using information processing skills were prevalent in all groups, with at least 50% (and as many as 100%) including something about this element in their response (Table 3.23). Aspects of the applying criteria element were mentioned in all of the group's responses, with Group A including it the most (83%), and Group G including it the least (13%). Aspects of the using existing knowledge element were mentioned in all but Group A's responses, with educators from Group U including themes related to this element in their response the most often (71%). Every group mentioned dispositions; however, the greatest difference in frequency of inclusion was between Group A (17%) and T (80%). Metacognition was the least included aspect that related to the framework, with no more than 20% of any group's respondents including it, and five groups not mentioning themes related to it at all (Groups G, I, A, O, T). Lastly, themes related to CT shifting information to understanding were mentioned in all but Group U's responses, with educators from Group T including themes related to this element in their response the most often of the groups (40%). Overall, the margin of differences between the inclusion of themes related to the framework were generally much greater using the SIMPROF groupings than the *a priori* groupings. Thus, showing the capacity for this multivariate method to provide deeper insights into the different ways that educators perceive CT.

Table 3.23 Themes from educator responses to Question 1a for SIMPROF groupings

	Factor	n*	framework element					enhancing component		
			purposeful querying	critical reflection	using information processing skills	applying criteria	using existing knowledge	dispositions	metacognition	information to understanding
important apriori groups	History	32	53%	59%	88%	50%	44%	50%	6%	28%
	Science	100	43%	25%	90%	46%	37%	46%	8%	22%
	High school	89	42%	26%	85%	30%	38%	49%	8%	22%
	Tertiary	43	53%	49%	98%	81%	40%	42%	7%	26%
	Science: High school	66	38%	18%	86%	30%	36%	50%	9%	20%
	Science: Tertiary	34	53%	38%	97%	76%	38%	38%	6%	26%
SIMPROF groupings	J	25	40%	8%	96%	20%	20%	56%	4%	8%
	V	20	30%	55%	90%	65%	45%	60%	10%	25%
	N	10	40%	20%	100%	40%	30%	30%	10%	30%
	O	10	50%	60%	90%	50%	60%	30%	20%	30%
	U	7	43%	43%	86%	57%	71%	43%	14%	0%
	G	8	38%	13%	50%	13%	50%	75%	0%	25%
	I	7	57%	43%	71%	29%	43%	43%	0%	29%
	A	6	50%	17%	100%	83%	0%	17%	0%	33%
	C	6	67%	17%	83%	50%	50%	67%	17%	33%
	P	4	25%	0%	100%	25%	50%	25%	0%	25%
	T	5	80%	100%	80%	80%	20%	80%	0%	40%
	*	24	54%	38%	96%	63%	42%	33%	8%	29%

*14 of the 146 participants from the study did not answer this question. The sample size for each group has been updated accordingly.

3.5 Discussion

This doctoral research into educator perspectives is the first cross-disciplinary, multi-institutional survey relating to CT in Australia. The core purpose of this survey in contributing to the broader thesis goals was to identify the current level of understanding that Australian science educators have about CT. It focussed on identifying potential gaps in educator knowledge, as well as examining perspectives about current teaching practice. This was achieved through comparing the perspectives of science and history educators (to determine whether there were discipline-based features evident in their views), and comparing differences between high school educators and tertiary educators (to see how variation differed among educators who have a set curriculum framework versus being able to generate their own educational outcomes). Because of its multi-faceted nature, CT is a challenging construct to define (Lai, 2011). As Mulnix noted, "what counts as CT seems to vary widely" (2012, p.464). This complicates educator decisions about how to target CT in the classroom. Yet it remains an essential attribute for the 21st century (ŽivkoviL·, 2016; Chan, Fong, Luk, and Ho, 2017). Therefore, it is imperative to keep assessing the understanding of educators and offer training and resources where needed so that the resulting standard of education and quality of Australian graduates remains competitive with other first-world nations.

3.5.1 Key findings

Since the specific objective of this component of work was to *investigate Australian educator perceptions about CT and its development and assessment,* one main key question, with four sub-questions were devised (see the Introduction or Chapter 3 Section 3.1 for a reminder of what they were). The key question was to determine whether Australian educator's views about CT differed depending on their discipline or education setting. In order to answer this question, the global profiles (comprising all 80 items) of the 146 educators, were compared using the PERMANOVA routine. Despite the conventional wisdom that there would be underlying differences between educators from different fields and settings, this study's experimental design, being a first of its kind, was able to test and identify for any differences between groups of educators across institutions and education settings. There was indeed a highly significant difference between the disciplines (science versus history) as well as education setting (overall high school versus overall tertiary; as well as science high school versus science tertiary). There are currently no other studies that have determined that high school educators think differently about CT to tertiary educators. Also, there are currently no other studies that have determined that science educators think about CT development and assessment differently to other educators (like history). However, there was one previous study that found that tertiary history teachers have a different conception of CT than tertiary physics teachers (Jones, 2006). Together this new study and Jones' study (2006) show that there are discipline-specific attributes related to CT and that a purely general approach is unlikely to teach students these things.

The first sub-question was to determine whether Australian educator's views about the nature of CT differed depending on their discipline or education setting. In order to answer this question, the conceptual sub-profiles (comprising of 11 items) of the 146 educators, were compared using the PERMANOVA routine. At the conceptual sub-profile level, there were no significant differences identified between any groups of interest. The ideas in the most commonly selected items concerned embedded CT development, CT involving knowledge, thinking skills and dispositions, and CT developing over time. These are all themes related to CT in the Australian Curriculum and TLO's (Jones et al., 2011; ACARA, 2011, 2013). This suggests that educators (or their employers/management who advise them) do use these documents for guidance about CT.

No study has previously compared conceptions of CT across high school and university teaching levels. However, it was thought that there could be a difference in perspectives since there are different policies ideas about CT in the two education settings, and there is also variation in the emphasis of CT as a learning outcome between the settings.

The similarity between the science and history educator conceptions, particularly at a tertiary level, is somewhat surprising, as previous studies have found conceptual differences. For example, Jones (2007) noted a difference in the conceptions of history versus economics educators from two Australian universities. In another article (Jones, 2009), Jones noted there are context-dependent attributes related to CT, and that the complexity of the rules applied during CT increases as students increase their discipline expertise (Jones, 2013). Bailin has also noted that science educators often conceptualise CT as skills and processes (2002), where history educators often conceptualise it in several ways including “as judgement…as scepticism…as a simple originality…as sensitive readings…as rationality…as an activist engagement with knowledge…and as self-reflexivity” (Moore, 2013, p.506). One possible reason that this study did not find these differences is sample size. Because of low response rates, educators from biology, chemistry and physics were all grouped under science. However, there is likely variation in the perspectives within these disciplines, which could have masked a clear discipline trend. There were also not very many history educators who responded. In addition, there was only one quantitative question relating to educator conceptions, so perhaps additional quantitative questions were needed to flesh out the key differences in conceptual ideas.

There was a supporting qualitative question concerning conceptions about CT (Appendix A: Survey Template Question 1a). These were analysed using the framework components first described in Section 2.2. Again, analysis by *a priori* groups revealed little difference in themes of the responses given by educators from different demographic backgrounds. However, when the SIMPROF findings were applied, the margin of differences between the inclusion of framework themes across the SIMPROF groupings were generally much greater. This shows the capacity for this multivariate method to provide deeper comprehensions of the different ways educators perceive CT. This question also provided valuable insight regarding the robustness of the decision concerning metacognition as an enhancing component rather than core element, since themes that aligned to critical reflection were mentioned more frequently than themes that aligned to metacognition. If there had there been more of an emphasis on self-reflection and self-evaluation in educator responses, there would have been a need to further explore, and perhaps reconsider, the role of metacognition as a core element in the ACT Framework.

The second sub-question was to determine whether Australian educator's views about the development of CT differed depending on their discipline or education setting. In order to answer this question, the development sub-profiles (comprising of 36 items) from the 146 educators were compared using the PERMANOVA routine. With regard to the development of CT, there was a highly significant difference between the disciplines (science versus history), as well as between the education settings (in the overall high school versus overall tertiary comparison and for the high school science versus tertiary science comparison). This shows that there are discipline-specific perspectives related to the development of CT. For example, when looking at perspectives on specific approaches, science educators indicated that the scientific method was the most effective of the listed approaches for developing CT, whereas history educators indicated that Socratic questioning and argument analysis were the more effective of the listed approaches. Educators also had the opportunity to list other development approaches and explain their perspectives, but few chose to do so.

The discipline-specific preference for the scientific method by science educators is not surprising. Schmaltz et al. (2017) describe how "scientific thinking provides students with the tools to distinguish good information from bad" (p.1). Science is thought of as an opportunity to challenge thinking to obtain "new knowledge…insights… and understanding" (Kaptan and Timurlenk, 2012, p.764). These ideas concerning new understanding are also core to the purpose of CT. The scientific approach also has similarities to CT because there are rules about what constitutes sound methodology and evidence, thus supplying another CT component (i.e. criteria for making critical judgements). However, Kaptan and Timurlenk (2012) have also cautioned that it is only when students get to practice as scientists and escape the "foundationalist emphasis on basic concepts" (p.764), that they experience sciences problem-solving potential. This means the challenge in helping students to become better critical thinkers through the scientific method, is to make sure the thinking and methods of science are more transparent to them as they gain foundational knowledge. To train the students to be skilled scientific problem solvers and competent critical thinkers, it is important that they experience and understand appropriate ways to question and test scientific knowledge. The need for a mix of content embedded and explicit problem-solving skills is supported by meta-analysis about CT development (Abrami et al., 2015). Further, as Bailin notes, "ultimately, judgment with respect to the application of principles is developed through an understanding of the practices which constitute critical thinking and the point of these practices" (2002, p.370).

The other main trend that emerged from the findings about CT development is that the majority of educators indicated they used an embedded approach. This was evident in both educator choices in Question 3 and in the responses to the qualitative follow-up question, where educators described approaches that generally relied on making assessments about CT development through asking questions and classroom/homework tasks that incorporated CT with content. This embedded approach is consistent with suggestions from the Australian Curriculum (Australian Curriculum, Assessment and Reporting Authority, 2013b) and TLO's (Jones et al., 2011). However, a meta-analysis by Abrami et al. (2015) has shown that an embedded approach, on its own, is the least effective way to develop CT. This represents an excellent opportunity to improve CT development in Australia, perhaps using a mixed approach, which is what Abrami et al. (2015) reported as the most effective approach for developing CT. However, it would also be worthwhile and useful to quantify the effects of current practice, to add to understanding about CT development. Contributions to this understanding are made through the case studies presented later in this thesis.

The third sub-question was to determine whether there were discipline-based or education-setting differences connected to Australian educator views about the methods and tools used for the assessment of CT. In order to answer this question, the assessment sub-profiles (comprising of 33 items) of the 146 respondents were compared using the PERMANOVA routine. There was a highly significant difference between the education setting groups (overall high school versus overall tertiary; as well as science high school versus science tertiary) in regard to the development of CT. For example, high school educators displayed a higher preference for self-developed versions of particular CT tools. Within the science only cohort, views differed on the type of CT test to use (i.e. general or discipline-specific), and whether or not the educator would be likely to self-develop it. This shows that there are education setting-based assessment approaches related to the development and assessment of CT. This had not been determined previously.

Despite these differences, the educators who were part of this doctoral research indicated they could identify CT in their students' work (with only 10% indicating they had difficulty, and 12% indicating they do not look for evidence). Most educators used an embedded assessment approach (~70%), but there was a difference in preferences concerning the source of the assessment tool, ranging from self-developing tools to using externally sourced instruments. Wiliam (2010) reviewed ideas around formative assessment and emphasised its importance because the plans and intentions of educators are not always what is experienced. Thus, if CT is an intended learning outcome, educators need to be careful to choose assessment tools that allow them to evaluate this with certainty. Educator responses to the survey revealed that there were many different procedures for assessing CT, ranging from CT tests to rubrics and course evaluation forms. Many of the most popular choices are tools that are either summative or left until the end of the learning period. This is problematic because it is too late to intervene if there is a problem. However, in the qualitative follow-up question (6a), many of the educators who responded indicated that they informally judge students CT abilities through classroom-based interactions with the student (i.e. through questions and discussion). These responses, in combination with some themes that emerged from Questions 3a-3c, confirmed trends from Question 5, where the majority of educators indicated they used embedded approach to CT assessment rather than assessing it in isolation. This tendency for embedded assessment also has the potential to be problematic if it means that CT has been tacked on as an afterthought, rather than strategically included in the assessment design. It not enough to think or expect that student CT abilities develop during college or university, there needs to be accountability (Roksa and Arum, 2011). The literature suggests that CT assessment needs to be grounded in a definition (Possin, 2008; Liu, Frankel and Roohr, 2014). However, the role of a definition in assessment was not a theme that emerged from educator responses. In fact, item C_1e, concerning standards and criteria, was one of the least selected statements from Question 1, and only 27.4% of educators contributed a response for the qualitative follow-up question for assessment (qualitative question listed in Appendix A: Question 6a). This is not overly surprising since a definition for CT is generally lacking in Australian policy, particularly for the final stages of high school and for university.

The fourth and final sub-question was to determine whether factors such as teaching experience, state (as a proxy for identifying state-based policies), or gender, helped explain differences in educator perceptions about CT. In order to answer this question, the global profile and sub-profiles for all 146 respondents were explored, per factor, using the PERMANOVA routine. Several pairwise comparisons returned highly significant results when comparing educators by state and education setting. For example, see Table 3.2. Of particular note is the difference between the high school educators from South Australia and Western Australia; and the high school educators from South Australia and Victoria. This difference between South Australian high school educators and the Western Australian and Victorian educators is not unexpected. Even though the Australian Curriculum puts forward a stance on CT for foundation to year 10 classrooms (ACARA, 2011), it does not apply to the majority of the Year 11-12 teachers sampled in this study, since it is only enacted for South Australian SACE (Government of South Australia, 2015). Since there is a difference, it would be interesting to investigate whether the differences in conceptions lead to differences in outcomes for CT development in future studies. If the South Australian students are found to have stronger development outcomes, this would send a strong message to policymakers to introduce nationally mandated policies related to CT for all schooling levels.

While not highly significantly different, there were some conceptual trends relating to state-gender and gender-teaching experience that emerged from the data set. For example, there were significant differences in the perspectives of male versus female from Victoria, and the perspectives of male versus female from NSW. As explained in the results section, there were also three differences related to gender-teaching experience. This warrants further investigation with a larger sample size since gender differences can be proxies for other trends. For example, there are "gendered patterns of academic choice in mathematics, science, and technology" (Yazilitas, Svensson, de Vries and Saharso, 2013, p.525) which in turn affects career options; there are also differences in educator conceptions which influence interactions between the educator and students (de Kraker-Pauw, van Wesel, Verwijmeren, Denessen, and Krabbendam, 2016). Thus, it makes sense that differences in educator perspectives about CT could be linked to factors the survey did not fully explore because of the sample size (such as perspectives within each science discipline).

For the global profile, 80 quantitative items were used to explore educator perceptions about the nature of CT, the development of CT and the assessment of CT. When analysing the survey in its entirety, this study did uncover some differences in science and history educator perspectives concerning the development of CT. Further, some differences in high school and tertiary educator perspectives emerged concerning the development and assessment of CT. Overall, exploration of the differences between educators revealed that there were some discipline-related and education setting related aspects to perspectives about CT, particularly concerning its development and assessment. In the global perspectives, there was also one highly significant and one significant finding, that related to education policy (as inferred through differences in state/territory-based perspectives). However, this study also identified clusters of respondents based on their global CT perceptions, who could not be identified by any one of the demographic factors tested in this work. This led to the conclusions that there are things that influence one's values and ideas about CT that extend beyond an educator's immediate setting and background. For example, in this 146-respondent cohort, it was possible to identify 11 core groupings which captured 80% of respondents (via SIMPROF routine), where each group had a particular set of characteristics that defined them as a group. While some groups were dominated by a particular demographic (such as Group G) comprised of only high school educators, other groups were a complete mix of demographics (such as Group U which had a 50:50 split of high school and tertiary educators and a 50:50 split of science educators and history educators). This is in line with above findings - meaning while there was a significant difference between science and history educators, there were also science educators who were more aligned with history educators (as evident in the ordination plots). This approach to classifying non-*a priori* groups is a new application in education research and perception studies, but it has been used in other disciplines like ecology. For example, in human microbiome projects, this technique was used to group people into 'ecotypes' based on their global bacterial profiles (see Szafranski et al., 2015). Whereas the more *a priori* style analysis is often used in perception studies; for example, a study on perceptions about marine protected zones among stakeholders (see Mangi and Austen, 2008). The beauty of the SIMPROF method in exploratory studies such as this perception survey is that it helped to clarify trends by cases that may have otherwise been explained away as outliers, or that may have remained as uncharacterised cases because they do not conform to an expected or known attribute.

3.5.2 Limitations of perception studies and suggestions about how to acquire more value and understanding from survey data

General limitations of educator perception studies include i) the number of participants (sample size); ii) participants coming from one institute or discipline; iii) only qualitative approach or a quantitative approach with limited assessable items; iv) when it is quantitative, only univariate analysis approaches have been used. Sample size can be hard to increase. Surveys are used extensively in education research (Desimone and Le Floch, 2004). Educators are bombarded with survey opportunities and need to decide their priorities for participation. Topic interest plays a role in survey participation (Groves, Presser, and Dipko, 2004); as does the mode of survey collection (Fan and Yan, 2010); gender has also been found to influence participation (Williams, 2008). Other factors such as "respondents' current moods; feelings of obligation, deference, and liking toward the survey… and perceptions that … participation is normative or represents a scarce opportunity to be counted" (Groves, Cialdini, and Couper, 1992, p.486) can also play a role. Web-based surveys also tend to receive lower response rates (Fan and Yan, 2010; Yetter and Capaccioli, 2010). These challenges are confounded when then needing to recruit sizeable numbers of participants from multiple disciplines/departments/institutions, particularly when there is no funding to incentivise participants for the use of their time.

Furthermore, a limitation of previous work is the underutilisation of the collected data. That is, either due to having rich qualitative data with limited statistical inference, or having quantitative data (like that produced from Likert-type scale or multiple-choice questions) that is often analysed using only univariate statistical approaches. However, once there are several items in a dataset, these set of dependent variables can be statistically analysed using a multivariate approach. Ewert and Sibthorp (2000) have suggested that when there are multiple variables, approaches such as an analysis of variance (ANOVA); multiple analysis of variance (MANOVA); or techniques looking for covariance - such as analysis-of-covariance (ANCOVA), can be applied. However, parametric approaches such as ANOVA and ANCOVA are limited by the total complexity they can manage (Anderson, 2001); they are not suited to skewed-data sets or varied standard deviations; and when appropriate for the data, can lead to Type 1 errors due to undertaking multiple pairwise comparisons. With larger participant sets, common multivariate techniques include Principal Component Analysis, factor analysis, structural equation modelling (Ewert and Sibthorp, 2000), or machine learning (Gabriel, Signolet and Westwell, 2017). Generally, the use of multivariate approaches in education studies is typically reserved for high-participant studies (for example, in the work of Kuh, Kinzie, Cruce, Shoup, and Gonyea, 2006; Gabriel et al., 2017; Windle, Haardörfer, Getachew, Shah, Payne, Pillai and Berg, 2018), where typically greater than 1000 participants have been recruited. However, there is a misconception that multivariate approaches can only be applied to such big datasets like those cited above. Yet the term *multivariate* explicitly refers to the number of variables in a dataset - in this case, the number of assessable items, rather than the number of participants (Clarke and Gorley, 2006). Thus, once a survey comprises several items irrespective of the number of participants, then multivariate analysis can be employed. For example, in the work of Shell (2001), the 43 Likert-type scale items could have been used to create a global profile of the 'challenges educator face when developing CT' for each of the 175 participants, where participants overall perspectives of CT could be analysed rather than considering each item separately.

Educators can find "using data from a big survey to improve student learning... hard" (Blaich and Wise, 2017, p.32). Generally, this means that discussions to determine the contextual application of the results need to be undertaken to establish the link between the finding and the "scholarship of [its] application" within a local setting (Blaich and Wise, 2017, p.33). Educators may find this profile approach a nice intermediary for big data approaches, qualitative approaches and univariate quantities approaches. With the capacity of the global profile approach to be able to handle smaller participant numbers, but multiple variables which can enable richer analysis, this approach offers a balance between the depth of qualitative data and the ease of collecting and collating quantitative data without the need for large participant numbers.

Of the 2500 people that the survey that was distributed to, 180 responded (~7%), of which 146 had complete quantitative response sets. While this only represents ~6% of the potential study population, those that did respond, self-selected themselves into a group that must have a particular interest to want to participate. While this participation rate could be seen as a limitation, it is an advantage because any differences observed likely represent true differences since it is comparing motivated respondents with motivated respondents (meaning the respondents' data are equivalent). Further, even though the number of responses is small relative to the size of the sampled population, it is larger than Moore's qualitative study (2013) on the conceptions of 17 history, philosophy and literary/cultural studies educators from a single Australian institute. It is also larger than, Jones (2006) study of 37 educators from economics, history, law, medicine and physics from two Australian universities. A recent Australian study by Danczak et al. (2017), had a higher respondent number (n=620 collectively across their target groups). This larger study consisted of a qualitative survey of chemistry students, educators and employers exploring the construct of CT. However, there are still limitations to this larger study, because the majority of the respondents were students or teaching assistants; the study was purely qualitative; it was from a single discipline, and respondents were mostly from a single institute. Therefore, while it provides a richer understanding of perspectives within chemistry at that institution, the broader applicability of the study is limited in its capacity to provide a general understanding of chemistry educator perspectives of CT. This study presented in this chapter overcomes these limitations in that it: has both quantitative and qualitative questions; the sample of respondents in this study represent multiple states and territories, and four disciplines (although due to the sample size all but one of these are grouped under the umbrella 'science'); and it uses multivariate approaches. It therefore represents a broader cross-section of perspectives across institutions throughout Australia, contributing an improved statistical approach for educators to consider.

The survey tool developed for this study revealed some limitations. These included survey length and question style. Concerning lessons learn about the question styles - should others consider using the educator perception survey tool developed for my study, it is worth noting that Stedman and Adams adapted Choy and Cheah's work when they developed their question set. This means that if future participants provide thoughtful responses, the results obtained using these qualitative survey questions should complement and would be comparable to responses to Stedman and Adams (2012) Likert-style statements. Alternatively, future studies may want to opt to use Likert-style statements to increase the chance of participation.

Concerning study length - engagement with the survey declined with each question – so keep them short. The number of questions in the survey had already been reduced from initial plans; however, to explore for both conceptions and perspectives of an educator's actions, the survey needed to investigate educator understanding of the nature of CT, as well as choices relating to the development and assessment of CT. Due to wanting to capture this overarching perspective of CT, but having a survey length that was not entirely off-putting, compromises had to be made. Questions focussing on CT development and assessment were prioritised, as they were the ones that would produce data to guide the case study choices. However, this choice was at the expense of clarity about educator conceptions of CT. In addition, the point spending question style used for the conception question (Question 1) was not as discerning as expected. It was structured to convey a range of theories about CT; however, the majority of respondents spent points on the same items. Further, in Question 1 the statement with the theme including embedded development was allocated at least one point by 80% of respondents, yet when asked about their development approach in isolation (Question 3), only 55% indicated they used embedded approaches. In addition, a greater percentage (38%) of respondents indicated they used explicit approaches than was evident from the point spending in Question 1 (7%). There are two possible interpretations for this. It could indicate that respondents have answered the survey inconsistently (due to either the framing of the statements – Fowley, 2009, or inaccurate reporting – Akbulut, 2015), or it could mean there is a discrepancy between educator's ideas about CT and the actions they take to help develop CT. The cognitive load of the question structure was quite high, so this could have led to confusion or inconsistencies (Fowler, 2009; Tourangeau, Rips and Rasinski, 2000). Especially given that this survey was distributed in an online environment, where the self-administered delivery mode means it is not possible to clarify unclear terminology and interpretation from the respondent. However, it is also possible that educators think and act differently when it comes to CT. Previous studies have found that educators think they are effective at developing CT, yet measures of student performance have not found this to be the case (Paul et al., 1997b). It follows that educators may perceive the nature of CT in a certain way may not be aligned to the actions towards their goal of developing CT. Irrespective of which, this finding further emphasises the need to gather additional information about educator conceptions of the nature of CT.

Future studies could seek to further explore educator conceptions about CT through a series of items, and then include one or two open-ended questions about assessment and development, since it is now established that there are differences by demographic factors such as discipline and education setting. Alternatively, studies could follow approaches such as Jones (2006) and Danczak et al. (2017) where interviews were used to seek for deeper understanding. Interviews were not used as a follow up for the findings from this study, as the broader goal of this research was to quantify CT performance outcomes. The perception survey was undertaken with the intent to highlight the types of educational experiences to explore and inform case study choices so that the quantitative data could be used to improve understanding about current educator practice. However, further verification of the findings through interviews and/or a revised survey could be undertaken in future studies to increase understanding of educator conceptions about CT and its development, as well as increase the confidence in and capacity for typing educators using the SIMPROF groupings.

3.5.3 Implications for developing CT

Through this study of Australian science educator perspectives about the nature of CT, the development of CT and the assessment of CT, it was determined that discipline and education setting help shape educator perspectives about CT. While there have been other smaller studies on the perspectives of Australian educators about CT, this study is an advancement for four main reasons. It had a greater number of participants from a greater range of disciplines and education settings, a greater number of survey items, and a more rigorous statistical approach for testing hypotheses.

It also confirmed expectations (or the conventional wisdom) that there would be differences in perspectives by demographic factors such as discipline and education setting, which in itself, has numerous implications for educators and policymakers alike. The statistical methods and the construction of global profiles contribute a novel application of multivariate analyses to perception survey data. By constructing perspectives from this global set of questions, the multivariate statistics allowed us to see that there were conceptual differences in the educators, just not by *a priori* groups. This indicates that there are educators who have similar ideas even though they are from different backgrounds, as well as educators who are working from their own conceptions and interpretations of CT literature rather than a more cohesive approach. This is important because there is a need for conceptual clarity (Green, Hammer and Stars, 2009) as the way educators view CT influences their curriculum design (Barrie, 2004). Existing literature on CT also shows that there are a vast number of perspectives on CT. Many education systems and institutions are still grappling with the complexity of CT, but the underlying issue is that lack a definition for CT means it is harder to bring accountability, as everyone could be doing their own thing. However, in this study, I have revealed similarities across educators. For example, the survey revealed that educators held four main ideas concerning the nature of CT, even though there is no policy-defined or commonly used definition of CT in Australia at a senior secondary or tertiary level (c.f. South Australian SACE).

The need for a clear and common definition really underpins all aspects of CT in education. As Ab Kadir (2017) suggests, “it is the teachers’ knowledge base of thinking that is a function of implementing a thinking curriculum” (p.81). Zohar and Schwartzer (2005) made four recommendations regarding the knowledge and approaches educators need to be able to “make the transition from traditional instruction that centers on transmission of information to instruction that sees the development of students’ higher-order thinking as one of its major, explicit goals” (p.1597). Among their recommendations, they argue that teachers need to understand CT and higher-order thinking skills things on a "cognitive level" and a "metacognitive level" (p.1597). They also argue that educators “need to implement a curriculum with higher-order thinking goals” and be able to “identify students’ reasoning difficulties” and be able to remedy them (p.1597); as well as engaging students in “the language of thinking” (p.1597). Previous studies exploring educator barriers to CT development may have revealed that there is a lack of confidence in their ability to incorporate CT. However, this study has shown that while there may be areas of uncertainty, there is a common thread in behaviours and perspectives that relate to discipline and education setting that both educators and policymakers can target for educational change.

These common areas can be used to advantage, by tracing these approaches back towards a relevant CT definition. The identification of this definition will help educators to be strategic in the development and assessment approaches by creating a boundary (definition/framework) for them to ground their pedagogy in. For example, when incorporating a CT definition into CT assessment, Australian educators have two main options. The first option is to map their current assessment strategies back to a specific definition of CT. The second option is to choose an external CT assessment, ideally a standardised one, which explicitly states which particular CT conception it aligns to, and then incorporate both the assessment tool and definition into their practice. If opting for the latter option, educators could choose a CT test such as the California Critical Thinking Skills Test (1990c; Insight Assessment, 2018), which is aligned to the Delphi Panel's definition of CT (Facione, 1990b). Alternatively, educators may prefer something with a more skills-based definition, like the Critical-thinking Assessment Test (CAT) (see Stein, Haynes, Redding, Ennis, and Cecil, 2007). The CAT has an explanation of the broad conception of CT it is based on in the training manual (Center for Assessment and Improvement of Learning, 2013), it also makes suggestions about the various experimental designs an educator might employ to study CT development. For descriptions of other CT tests which are available, see Liu et al. (2014) and Ennis and Chattin (2015). Rather than using a formal CT assessment test, educators may choose to test for CT skills and abilities on a more individualised basis, in which case they may wish to self-develop formative assessments pertaining to particular CT skills as relevant to their discipline and preferred CT definition. If using this approach, educators might also like to consider the work of Bensley and Murtagh (2012) who outline some guidelines for a "scientific approach" to CT assessment (p.5) – however, their suggestions are also broadly applicable to other disciplines. Otherwise, to create a more cohesive and universal approach, the policymakers in the Australian education system could identify the CT conception they want Australian educators to incorporate into their practice, and could then establish and offer relevant training and resources in line with this perspective. However, some caution should be taken with a more globalised approach because "real-world assessment problems resist one-size-fits-all solutions" (Wright, Goldwasser, Jacobson and Dakes, 2017, p.45). Therefore, a conception of CT that is sensitive and adaptable to different context (such as the ACT Framework) should be used to help standardise CT in Australian education.

3.6 Conclusions

This new study supplies a survey tool, which, while not free from limitations concerning length and clarity, captures more aspects of perspectives about CT than any other existing tool. If some amendments were made to the conceptual question, this tool could be used as a starting point for site-level explorations, or for creating a baseline in education systems where educator perspectives on CT have not yet been characterised. The chapter also demonstrated the capacity for my Adaptive Critical Thinking Framework to enhance understanding about educator perspective about CT. However, the main contribution of this study to education is the multivariate statistical approach employed (used to analyse the global profile in its entirety), because it made it possible to determine that the underlying differences in educator views could not be fully explained through the *a priori* features measured. Therefore, while there were some significant findings where education setting and discipline seemed to shape choices relating to the development and assessment of CT, there were other factors that also influenced an educator's approach to CT. This suggests that the multivariate analysis approach, and its ability to generate profiles from perspectives, is more powerful for discerning trends among educators than seeking for these differences per survey item (or item sets), as is typically done in education research. In the context of this study, the SIMPROF groupings provided deeper insights into CT approaches, which policymakers can then factor into curriculum design. It highlights that science educators do not always use the same approach nor do history educators, which means that even within a discipline, there is no one-size-fits-all approach. It also provides insight into student preparation for university. Knowing that there are education-setting differences means that the tertiary educators can plan to be more explicit about defining the context and expectations around the use of CT to help students reach an appropriate university standard sooner. It also highlights the opportunity for bridging a gap in the transition to university, whether it be changing something at the high school end, or making it a core component of a student's first year at university.

Ultimately, the approaches used throughout this study have demonstrated that differences in educator perspectives about CT are not always clear-cut. So, while this holistic approach was able to show that there are some education setting-related and discipline-related perspectives that shape choices relating to the development and assessment of CT, it also revealed that the underlying differences in educator views could not be fully explained through the *a priori* demographic features measured. However, the real power of this global profile is that it revealed that when it comes to ideas about CT in Australia, there are groups of like-minded educators from different disciplines and educations settings. This highlights the need for greater population studies and further work to establish more understanding about these educator groups, to determine how best to support them to maximise CT outcomes in their classrooms.

This chapter has completed the exploration of the conception of CT in this dissertation research. It has provided original insights into conceptions of CT from Australian high school and tertiary science and history classrooms, and has also demonstrated the power of multivariate analysis to supply deeper insights into survey findings. Through the exploration of themes around CT development and assessment, this chapter has provided first insights into the next lens this doctoral research examines CT through. The next three chapters continue this journey into classrooms, interrogating CT development outcomes using a novel multi-tool assessment approach.

PRIMER 3

The last chapter was a statistics-heavy investigation of educator perspectives. It was interdisciplinary in its application of scientific methodologies with regard to survey analysis, but walks the border of transdisciplinarity because of the way in which I present the information, and in the use of science and social science methods to answer and education-related questions. As evident in the lack of literature, social scientists would be unlikely to use sample similarity matrices in their analyses. However, the stakeholders also helped frame the presentation of the results, in that a scientist would generally only present and discuss statistically significant results rather than the results as a whole. But given the exploratory nature of the study, I took the approach of presenting all the findings – statistically significant or not, whether they are later found to be supported, challenged or disproven. This is the nature of knowledge generation.

The study produced a number of important and original contributions of knowledge. It confirmed expectations about differences in educator perspectives when summoning CT. These distinctions were determined by factors such as discipline and education setting. Of particular value for educators and policymakers was the finding that different disciplines use different pedagogical approaches. Therefore, the best outcomes for professional development training should have a disciplinary focus. However, conceptual collaboration about CT should not be limited to discipline, as the approach of creating and exploring global profiles revealed like-minded thinkers from different demographic backgrounds. This study also demonstrated the value of the framework, by showing how the framework helps by further clarifying the emergent educator profiles by their perspective on CT. In particular, the framework in conjunction with the statistically significant SIMPROF groups also brought more clarity to the thematic analysis of educator responses to Question 1a helping to demonstrate that there were key differences in how some groups of educators conceptualise CT compared to others that were masked by the *a prior* groups. Overall, this component of the research contributed both novel knowledge and methods. Given the sample size, perhaps the more significant of these contributions is the application of multivariate analyses to perception survey data showing how these techniques can be used to obtain more value (deeper insights) out of survey data.

Transdisciplinary research results in an amalgamation of methods and knowledge. In this first study, understanding of educator perspectives on CT was enhanced by scientific analysis methods and the ACT Framework elements which has a number of implications for curriculum development and training educators about CT. But over the next few chapters the blend of qualitative and quantitative knowledge that emerges helped to reconfigure my perspective on what constitutes valuable knowledge and acceptable evidence at the boundary of education research. This blending of methods between science, curriculum development, and critical thinking has resulted in a harmonious mutualism, where understanding about CT becomes slightly more clarified at the points where each field intersects. This also highlights that the knowledge generated through transdisciplinary research is not one-directional, nor its application bound to one context.

Scientist and indeed education institutions have "traditionally valued objective knowledge as the highest form, whereas teacher research recognises that all knowledge is subjective" (Abell, 2005, p.293). Having begun this journey as a scientist that was interested in how science is communicated and taught, it should not be surprising that I place an emphasis on quantitative evidence. But as will become evident in the assessment tools I have chosen and the results I present in the following three chapters, I have a new appreciation for the capacity of qualitative findings to add richness to quantitative results. This was an uncomfortable yet conscious decision; the scientist in me finds safety in numbers and statistical significance. But as recently reported in Nature, it is detrimental to use statistical significance (P-values) as the gatekeeper of important knowledge (Amrhein, Greenland and McShane, 2019). Further, in this journey, I have learnt that education research is messy because humans are much harder to control. Classroom environments present challenges scientists usually would not face - you cannot make students retake a test, or expose them to trial after trial after trial. However, I have still drawn on my scientific understanding of good research practice. A scientist would never go into an uncontrolled environment and take one kind of measurement to explain an effect in that environment – even in a behavioural study. For example, when trying to understand behaviour, ecologists examine both the inter-species and intra-species relationships, as well as dynamics between their target species and their environment. Scientists often have the benefit of multiple study sites where they can gather data from, or can return to over a period of time; whereas classroom research is a one-time event (Abell, 2005).

While I know that classroom research was less controllable and not repeatable (in an experimental sense), my biggest qualm remained. For too long, education research has supported my specific methodological choices – be it qualitative or quantitative. Yet, when scientists cannot control variables, they try and collect as much data as they can so they determine which factors might be causing an effect. This is a frustration for many of my science colleagues, especially those who regularly invest time in developing and modifying their curriculum and pedagogy.

These next case studies use an observation tool, a perception survey and a performance measure to gain deeper insights into CT development in the classroom ecosystem. The case-study components of my research dance between Scholarship of Teaching and Learning and Discipline-Based Education Research purposes, thus making it a transdisciplinary form of what are already interdisciplinary research processes. This research is certainly applicable to both specific and broader STEM classrooms, however, the overall understanding about the types of experiences which lead to CT development are applicable throughout education (not just science).

Chapter 4

Exploring Critical Thinking Development in Chemistry

Education is not the learning of facts, but the training of the mind to think.

Albert Einstein, date unknown

4.1 Introduction and chapter outline

There is much literature on CT development, there is a paucity of information conveying the experiences and perceptions of Australian science educators. For example, since the publication of Facione (1991) and extending to the end of 2018, there are between 739 - 1400 publications that include the terms 'critical thinking' and 'development' in their title (according to Scopus and Google scholar, respectively). Of the 739 articles found on Scopus, there are just 57 studies from Australia, with only handful about science classrooms. Thus, there is a need to further explore CT development in science in Australia so that future changes to (and recommendations from) Australian policy can be informed by contextually relevant findings.

In Chapter 3, I presented findings about the approaches Australian educators reportedly use to develop CT skills. Embedded techniques were a consistent theme for how these science educators approached both CT development and assessment. Educators also conveyed perceptions of actions that are effective for developing CT. The general impression held was that their approaches were having a positive effect on CT development, even though most educators were not doing anything specific to assess this change. In fact, when asked about their specific actions, few responses included using a standardised assessment instrument, particularly in a tertiary environment. Additionally, there is minimal empirical understanding about the effect of using embedded approaches in the literature for domain specific courses such as science and the arts (Tiruneh et al., 2016). This next study addresses this knowledge gap through a structured quantitative analysis of science classroom experiences. It starts to answer the second broad Research Question: *What teaching approaches are currently employed by science educators in order to develop critical thinking abilities, and which are most effective?*

This chapter explores CT development in chemistry. Previous research on chemistry and CT has revealed that studying chemistry can lead to changes in CT dispositions (Qing, Ni and Hong, 2010); as well as changes in student CT performance (Espinosa, Monterola and Punzalan, 2013; Fensham and Bellocchi, 2013; Kim, Sharma, Land, and Furlong, 2013). However, there is still quite a knowledge gap about how studying chemistry lead to these changes. Approaches involving interventions such as games (Henderson, 2010) and writing activities (Oliver-Hoyo, 2003) have been investigated. Since the commencement of this research there has also been more exploration of writing (Gupta, Burke, Mehta and Greenbowe, 2015; Stephenson and Sadler-Mcknight, 2016), scientific literacy (Miller and Czegan, 2016) as well as an investigation into the effect of problem–based learning (Cowden and Santiago, 2016). While all have led to positive outcomes in CT development, the findings and implications were context-specific. In addition, none of these studies were conducted within Australia. A more generalisable finding came from Jacob's study (2004), which concluded that chemists need to be more explicitly trained in CT so they can evaluate statements and engaging in chemical reasoning, as well as being capable of applying their scientific evaluations to other knowledge domains. The suggestion of explicit CT training in discipline-specific contexts aligns with findings about mixed approaches, leading to the greatest CT outcomes in college (Abrami et al., 2015). Given my survey findings, in addition to the fact that suggestions of embedding CT dominate Australian policy, it is important to explore the outcomes of Australian classrooms before embarking on widespread policy change and educator training. Thus, this study is important in starting to remedy the gap in understanding how studying chemistry in Australia, where predominantly embedded approaches are used, might develop CT skills. This chapter also demonstrates the capacity for my Adaptive Critical Thinking (ACT) Framework to assist with the targeting of specific CT abilities to inform pedagogical improvements.

However, before exploring the effect of chemistry on CT, this chapter first considers some of the theory and existing findings around CT development and assessment that were not covered in the literature review. It draws on current understanding about CT development, particularly in tertiary settings, and explores the predicaments educators face about selecting CT assessments. It discusses why the prevalence of embedded approaches in Australia is problematic. It also establishes the need for clearer assessment of CT within Australia, followed by the case study itself which models a way to assess CT development.

4.2 Reflection of existing research on CT development and assessment

Numerous meta-analyses have revealed a range of approaches are used for developing and assessing CT (Allen et al., 1999; Ortiz, 2007; Niu et al., 2013; Abrami et al., 2015; Huber and Kuncel, 2016). Therefore, the sheer number of perspectives and approaches, in addition to the complex nature of CT, can make it difficult for educators to decide which theories and strategies to integrate into their pedagogy and practice. The learning environment has been found to influence student CT development (van der Zanden, Denessen, Cillessen and Meijer, 2018b; Abrami et al., 2008, 2015). Robert Ennis, one of the early theorists and contributors to Facione's Delphi panel has suggested there are 5 main typologies relating to teaching for CT (Ennis, 1989). These typologies are used to define the extent to which domain-specific knowledge and content play a role in CT development (return to Chapter 1 Section 1.3.2 for more information). Successful CT development has been measured in all typology formats (Abrami et al., 2015). In addition to their being different ways of incorporating CT into the curriculum (explicit versus embedded CT terminology and tasks), there are various pedagogical approaches which can be employed to develop CT. Ennis alone has proposed 21 strategies he has experienced as effective for developing CT (Ennis, 2013). Increased emphasis on approaches involving inquiry-based learning has added another layer to classroom experiences and learning outcomes, scaling the degree to which the learning environment is directed by the teacher or the student (National Research Council, 2003; Beck, Butler and Burke da Silva, 2014). While Abrami et al. (2015) found a mixed teaching approach, and tactics such as mentoring, dialogue, and authentic instruction had the strongest effects on CT development, there is still a lot to be understood about CT development.

Across two longitudinal studies, Pascarella and Terenzi (1991, 2005) have consistently found that tertiary education experiences have a positive effect on student CT development. But achieving a guaranteed CT development outcome is not quite as simple as attending college or university. In fact, the magnitude of the effect was smaller in Pascarella and Terenzi's more recent study (2005). Findings from the creators of the Critical-thinking Assessment Test (CAT) have revealed that a 4-year American university experience, without specific CT interventions, increases CT performance by an average of 25% (Harris, Stein, Haynes, Lisic, and Leming, 2014). Yet with interventions, they have also found that >24% improvement can be achieved in just a single semester (Harris et al., 2014). There is also evidence to show that not all degrees are equal (Niu et al., 2013; Huber and Kuncel, 2016). For example, Niu et al. (2013) identified that the development of CT skills in social science and health profession students was less than the development observed in students from other disciplines. Similarly, Karabulut (2015) also identified issues with the development of CT in social science students. Together these findings suggest that certain ways of approaching CT development are more effective than others.

In Chapter 3, Australian educators revealed a tendency towards using embedded approaches to teach for CT. This finding was not surprising since it is encouraged through Australian policy and curriculum, yet it is a tenuous policy position given the evidence in the literature. For example, Abrami et al. (2008, 2015) have consistently found that of the four most common ways to approach CT development, a purely embedded (immersive) approach produced the smallest effect on CT development. There are three main issues with embedding CT. The first is that specific CT learning outcomes may not be clear to students, especially if CT is not mentioned explicitly as a learning outcome. It also makes it harder for an educator to reflect on achievements specifically relating to CT, if CT is embedded among other outcomes within assessment tasks. A stance of embedding CT also makes it harder to set parameters for accountability across institutions. This same difficulty can also potentially arise within an institution, as every educator could have a unique approach to CT assessment (which could include poor quality measures). In their study of bioscience in Australia, Elliot (2010a) found that assessment of learning was mostly based on student perceptions. Elliot expressed that most learning evaluations ignored other forms of evaluation that Kirpatrick (1975) considers important for meaningful evaluation (such as measuring changes in student knowledge, behaviours over time as well as overall education experience). These low-level evaluation processes are a consequence of the old Australian Universities Quality Agency (AUQA) audit procedures, only requiring Australian universities to show how undergraduate degrees incorporate and embed graduate attributes (Barrie, Hughes, and Smith, 2009). However, changes to quality assurance requirements and government policies mean that the Tertiary Education Quality and Standards Agency (AUQA's replacement) is expected to start policing attribute development soon (Donleavy, 2012; Oliver and Jorre de St Jorre, 2018). To date, CT has predominately been incorporated into courses through basic alignment of learning outcomes to graduate attributes with little accountability to the actual outcomes.

The need to build in assessment for graduate attributes including CT, as well as explore the development of CT in education settings, has been reiterated by Stedman and Adams (2012). They suggest that "the way material is presented has a large effect on whether or not critical thinking takes place" (Stedman and Adams, 2012, p.9). The Delphi panel (Facione, 1990a) made some specific recommendations about assessing CT. They suggested that "assessment should occur frequently" (Facione, 1990a, p.17), and "there should be minimum proficiency expectations" (Facione, 1990a, p.16). The Panel also advised that CT assessment should be made explicit to reinforce its worth to students; and "different kinds of instruments should be employed" (Facione, 1990a, p.17). They also recommended that assessment strategies should "be guided by a holistic conception of what it means to be a good critical thinker" and avoid the trap of assessing "the more readily targeted" skill-based components (Facione, 1990a, p.4). Yet, actually implementing these assessment goals has been problematic, even with Facione's (1990c) own CT assessment tool. CT assessment has been approached from a number of directions, which is not surprising given the vast array of definitions and pedagogies. There are assessments which focus on skills and others which focus on dispositions[6] or perceptions.[7] There are approaches which measure CT exclusively (either generally[8], or using discipline-specific questions[9]), and there are others which use more of an embedded approach (either formative or summative[10]). This array of tests presents a further challenge to an educator's decision making about CT.

General CT skills tests are commonly used in the USA (Adams, Whitlow, Stover and Johnson, 1996). There are approximately 59 validated tests available. Commentary and summaries about these instruments can be found in Facione (1990b); Stein, Haynes and Unterstein (2003); Possin (2008); Ennis and Chattin (2015); Aloisi and Callaghan (2018); with some additional information in Arter and Salmon, 1987; Renaud and Murray, 2008; Ku, 2009; Saadati, Tarmizi, and Bayat, 2010; Csapó, Ainley, Bennett, Latour and Law, 2012; Sosu, 2012; Tremblay, Lalancette and Roseveare, 2012; and Benjamin et al., 2013. Not all of these instruments focus exclusively on CT, and some only focus on exclusive aspects of CT. The skills they measure vary, ranging from logic, thinking skills to dispositions, so educators should be thoughtful about the learning opportunities and learning outcomes offered in their courses when picking an instrument (Csapó et al., 2012). Test choice is highly important. In fact, Hatcher (2013) trialled three different general standardised CT tests and found that the results varied tremendously, even though the syllabus (and other course components) remained fairly consistent throughout the study. Upon further reflection of the results, Hatcher (2013) concluded the Ennis Weir Critical Thinking Essay Test (EWCTET) was the better gauge of his students' CT abilities compared to the multiple-choice question tests he trialled. This was because Hatcher felt the EWCTET had the best resemblance to both real-life scenarios and the experiences offered in his course. However, it is not just about aligning the test and learning outcomes to assess the relevant skills. Test choice is also important because certain question formats have been found to influence the student outcome. Hyytinen, Nissinen, Ursin, Toom, and Lindblom-Ylänne (2015) found that the way students' process information can affect

---

[6] Tests based on dispositions include Halpern (2010) and Butler et al. (2012).

[7] Tests including perceptions include Tapper (2004) and Choy and Cheah (2009).

[8] Some examples of general CT skills tests include Watson and Glaser (1980); Ennis, Millman and Tomko (1985); and Facione (1990c).

[9] For discipline-specific CT skills tests, some examples include ACT CAAP Operations (1989); OECD (2008); and Benjamin et al. (2013).

[10] Some examples of formative and summative CT assessments include Friedlan (1995); Polomba and Banta (1999); and Elliot et al. (2010a).

their performance on multiple choice questions versus constructed response questions. In Hyytinen et al.'s study (2015), 10% of students had completely opposite results on the two general CT tests they completed, even though the skills the tests measured were the same. Hyytinen et al. (2015) and Hatcher's (2013) findings clearly demonstrate that results can vary depending on the test used and using a standardised CT test does not guarantee that it is a comprehensive or relevant assessment of CT in any given classroom. This highlights the challenges educators face when trying to decide between existing general CT assessment instruments. Walsh and Seldomridge (2006) caution against using general instruments to evaluate thinking skills. Their impression is that general CT assessments are "unlikely to help improve curricula, instruction, or professional practice" (p.161) because the questions are not formatted to capture discipline-specific nuances that demonstrate appropriate CT for a given context. These findings and warnings about CT assessment reinforce the importance of the alignment of assessments to intended learning outcomes and ideas about CT, especially if the findings are intended to be used to inform future changes to a course. So even though general CT tests offer a consistent and validated approach for exploring CT development across courses and institutions, obtaining context-specific classroom evidence may be easier to achieve with a discipline-specific instrument.

Educators must include an array of materials and references in their courses, in a reconfiguration of the proliferation of information platforms. Therefore, adding in extra assessments which have not been found to provide meaningful data would be counterproductive. Yet finding discipline-specific tests, especially standardised ones, is a real challenge. In fact, Ennis and Chatting (2015) only describe four in their annotated list. Elliott, Boin, Irving, Johnson, and Galea (2010a) expressed that there are few examples of tertiary-level science specific CT tests, and there are similar issues in other disciplines, such as nursing (Adams et al., 1996). As a solution to the lack of available science-specific CT assessments, Elliott et al. (2010a) suggested that a way to measure CT and inquiry skills is to use a pre-test post-test design in which students are required to "design a scientific inquiry... [and then] extrapolate the information to a new situation" (p.64). Elliott et al. (2010a) suggest using this approach as a way of monitoring student progress. However, they do not explain what constitutes a sound response that is demonstrative of tertiary level CT. The ideas Elliott et al. (2010a) present are insightful, yet the assessment design lacks the rigour scientists expect when producing meaningful and comparable results. As a consequence, it does not address the issue of there being very few CT tests for tertiary science courses. However, there are some newer assessments that could meet this need, including the Collegiate Learning Assessment (CLA), the CLA+ and the Critical-thinking Assessment Test (CAT).

The CLA and CAT have open-ended questions that have been designed to evaluate CT in a more holistic way than a multiple-choice sub-scaled CT assessment (Klein, Benjamin, Shavelson and Bolus, 2007; Nusche, 2008; Lai, 2011a; Tremblay et al., 2012; Benjamin et al., 2013). The CAT is technically a general CT test because the information contained within each question-set was constructed to be sufficient to enable test-takers to answer the questions without further input (Stein et al., 2007). However, the questions have a science-feel because it was developed with input from American science educators. As such, the CAT resembles the closest version of a standardised science-specific CT assessment that is currently available for widespread use. The CLA is also a general test; however, there are ways to modify it to suit discipline-specific needs. For example, a modified subset of the CLA was used by the OECD's Assessment of Higher Education Learning Outcomes (AHELO) in an international study of generic skill development in higher education (Tremblay et al., 2012). However, this AHELO instrument is not available for general use. In addition, the CLA component of the tool was ultimately found to lack applicability across international contexts (OECD, 2013a). The CLA was also criticised for only showing if CT develops, meaning it was not useful for pinpointing strengths and weaknesses in CT skills (Benjamin et al., 2016). Consequently, an updated version was developed (Benjamin et al., 2016). However, the validity of the CLA+ has also been criticised on the basis of poor clarity around the aspects of CT it measures, the adequacy of the scoring system, and its capacity to be used for student-level decisions (Aloisi and Callaghan, 2018). Although not currently reported in the literature, the CAT could also be subject to a number of these flaws. However, the main benefits of the CAT over the CLA is that the test creators involved their intended audience in the test development as well as the test scoring process so that test users can have better insight into the results they get back (Stein et al., 2003). Further information about the CAT instrument is provided in Section 4.4.1, with a discussion of the lessons learnt in Chapter 6: (Section 6.4.3).

Despite the scale of literature, the findings have yet to truly clarify CT development for educators. There is some understanding of the broad techniques they should employ, but there is still ambiguity about some of the context-specific changes they might make to accomplish stronger the outcomes from their classrooms. However, to achieve these goals, the real challenge lies in the assessment of CT. There are number of issues relevant to most CT assessment tools, the implications of which are discussed in the conclusion: from concept to classroom. While newer, open-ended assessment instruments are a somewhat better fit for the criteria outlined by the Delphi Panel, there is still more work to be done. Especially in Australia where there are not many publications which explore CT development in science. The case studies presented over the following chapters attempt to address some of these issues by modelling a multi-tool CT assessment process used to increase the understanding of CT development in Australian tertiary science classrooms. The aims and methods for these studies are presented next, followed by the findings from the first case-study (chemistry).

4.3 Case study objective, aims and expectations

Science education is a growing field that incorporates methods such as education evaluation, scholarly teaching research, the Scholarship of Teaching and Learning approach, and Discipline Based Education Research (Singer et al., 2012; Dolan et al., 2017). The first three of these approaches tend to be focussed at a localised level, seeking to gather evidence to inform action or increase understanding about the impact of a course or program. This kind of research may or may not have broader applicability; however, of the three, Scholarship of Teaching and Learning (SoTL) is the most likely to lead to have broader practice implications because the research is generally made available for peer review (Shipley et al., 2017; Dolan et al., 2017). In contrast, the fourth approach, Discipline Based Education Research (DBER), functions to explore science education practice for the purpose of uncovering more "generalisable and mechanistic understanding about educational processes and their effects" (Dolan et al., 2017, p.3). Both SOTL and DBER techniques have influenced the way these case studies are conducted.

There are already a number of studies which explore CT development through CT performance, or through perceptions of student learning. However, my case studies are unique in considering classroom opportunities in addition to perception and performance changes. They bring together a series of tools, including the ACT Framework (Chapter 2), to demonstrate a way that existing courses could be evaluated without the need for extensive training. This collective approach provides the diagnostic ability to enable strategic targeting of specific learning outcomes and CT skills in future versions of the course. The first two case study chapters (5 and 6) are structured like SoTL research, focussing on gaining understanding relating to a specific course. However, the purpose of these individual studies was always broader and more in line with the other main science education research process (DBER). This is evident in the comparison of case-studies chapter (Chapter 6), which explores trends across the data to generate broader understanding about the effects of science teaching on CT development.

Henderson et al. (2017) notes that the DBER approach could benefit many fields, but its uses thus far have been limited to science. My case studies reveal the potential for wider use of the DBER method, because even though they examine science courses, it shifts science out of the primary focus of the DBER method, instead making it a lens to explore critical thinking development, in order to configure space for generating broader understanding about CT. A key aspect of DBER research is that the discipline sets the priority for the research - thus the disciplinary emphasis guiding the SOTL and DBER methodology choices stem from literature on CT, with the research questions focussing on CT development in science (note: science was the natural choice of focus given my background; however, the approach I have taken can be adopted in any field provided the assessment tools remain relevant to capture classroom opportunities offered).

The information from these case studies address research aim 3: *to explore how Australian tertiary science educators at Flinders University are developing CT and determine the effect of their approaches using a case-study method that could also serve as a general model for CT assessment.* This aim comprises of two main objectives:

1. To determine the effect of science courses on the students' CT performance and the perceptions of their CT abilities.
2. To demonstrate how different teaching approaches can lead to different outcomes in CT development.

The first objective, concerning the effects of two science courses, biology and society, and chemistry, on tertiary student CT development is addressed over the next two chapters. Meta-analyses have shown that certain features of a particular course (rather than a course in its entirety) can produce the biggest improvements in CT development (Abrami et al., 2008, 2015). Consequently, these courses were chosen to explore the different types of actions and effects that might be experienced in science classrooms. Each study compares the change in CT performance across a semester, in addition to changes in students' perceptions of their CT abilities. Classroom observations were also made to help capture the types of experiences, in attempt to relate them to any perception and performance changes.

The first case study explores the effect of a pure science course (chemistry). This course had an embedded CT approach and includes many opportunities for scientific inquiry throughout the course. Methods used in scientific inquiry have many parallels to CT processes (Lederman, 2008). It is thought that science, logic and the scientific method "provide us with models that we can attempt to…emulate in our thinking" (Lipman, 1987, p.5). Münnix (2018) notes the similarities between conducting philosophical thought experiments and science experiments, particularly in terms of the process of questioning and hypothesis generation. There is ongoing belief that science helps "students to develop their higher order thinking skills to enable them to face the challenges of daily life." (Saido, Siraj, Nordin, and Al Almedy, 2015, p.13). But there is a lack of empirically-based understanding about the effect of domain-specific learning environments, which embedded CT (Tirenuh et al., 2016). However, science educators from the Australian perception survey (Chapter 3) also indicated they thought that the scientific method was an effective way to develop CT skills. This course provided an opportunity to quantify these beliefs and investigate whether participation in general scientific approaches in an embedded CT environment leads to measurable changes in CT development.

For the second case-study, a science and society course (biology and society), rather than a pure science course, was chosen. Science and society courses enable students to explore how scientific issues fit into the world around them, which helps student to contextualise their learning (Llopart and Esteban-Guitart, 2017). In the USA, these kinds of courses have been shown to have a positive effect on CT development (Rose, Gillespie, Rowe, Primm and Shannon, 2012; Gottesman and Hoskins, 2013). However, educators can have misconceptions about what aspects of their pedagogy lead to the classroom outcomes they observe. For example, Ortiz (2007) tested the assumptions that taking philosophy would result in CT development. However, Ortiz actually found out that studying CT courses and other courses that included training in argument mapping (not necessarily a philosophy course) produced the strongest positive effects. This second course choice enabled exploration of the assumption that explicit/infusion CT instruction involving critically analysing themes and arguments around science in society will produce a positive effect on CT development. This course was studied twice, with the second iteration including an intervention.

The combinations of the two different courses also provided the opportunity to explore which types of experiences have a stronger effect on CT development. This relates to the second objective of aim 3 - *determining which classroom approaches had the greatest effect on CT* *development*. This aim was achieved by comparing the set of observations, perceptions and CT performance results from each science course, to determine which course had the largest effect on CT development. Findings related to this aim are addressed in Chapter 6 (Comparing CT development in science).

As these investigations are exploratory studies, to determine a suitable hypothesis, course coordinators were first questioned to find out whether they thought their course developed CT. Both course coordinators thought their courses did, therefore the hypothesis for each case study was structured with the expectation of a positive effect on CT development rather than assuming a null effect. In these studies, a positive effect is measured by a positive change in performance and perception. Since the student perception survey responses were on a Likert-scale, if the courses were perceived by the students to be effective at helping the students develop these skills this meant there should be a shift towards the strongly agree category across the semester. But what should the magnitude of the CT performance be?

Wasserstein, Schirm, and Lazar (2019) suggest the use of context, prior experience to determine how to thoughtfully interpret a statistical result. Brownstein, Louis, O'Hagan, and Pendergast (2019) claim that there is subjectivity in scientific inquiry, despite the fact that objectivity is the goal. Brownstein et al. (2019) also suggested not to under-estimate the role of an expert. In this case the expert on this tool are the creators of CT assessment used in these case studies – as the developed the instrument, and also have a large data set from which to determine the conditions for an effect versus a non-effect. The creators of the CAT have suggested that a 26% improvement (or 4-point score improvement) occurs across a 4-year college degree with no targeted CT interventions (Harris et al., 2014). The test creators do not have sufficient data to determine a change at a semester-level, but assuming that CT development is linear across the degree this roughly equates to 0.5-point improvement per semester. This means the score change should be greater than 0.5 per semester if something about the approach or interventions used in any of the case-study courses is improving CT development. However, this is one of the major limitations in this study, because the literature indicates that linearity is unlikely. For example, findings from Roohr, Liu and Liu (2017) revealed that CT skills did not significantly increase until completing four of five years of college. Similarly, Arum and Roksa (2014) also found that learning gains were greater in the final two years of college. However, in the context of the first-year courses explored in this dissertation, this assumption of linearity will result in an underestimation of the change and therefore is a conservative approach to explore the effects of the courses.

4.4 Case study methods, tools and rationales

In this section, the methods and tools applied in the case studies are outlined with a brief description of the justification based on Facione recommendations, and other published literature. For example, Facione (1990b) recommended using multiple instruments to gain a holistic conception of CT development. A multi-tool approach was important for my study because no existing assessments include all the facets of CT. In addition, Kirkpatrick (1996) has suggested that meaningful evaluation should include gauging reactions, exploring the learning (knowledge, skills and/or attitudes), investigating the transfer of the learning (behavioural changes) and identifying and (if possible) measuring the broader implications of the training. At an intuitional level the process to create change can be quite expensive. However, the methods employed in this case study have tried to balance the need for multiple measures, at a low resource cost so that the approach is accessible to educators at a classroom level. The case study process itself consisted of making classroom observations of educator and student actions, using surveys to gather educator perceptions of students and student perception of themselves, as well as collecting and conducting student performance assessments (through both coursework and specific CT assessment). Descriptions of the tool choices are covered in the next section (see Table 4.2), but in short, the rationale behind the tool choices are as follows: -

- The CT performance measure (Critical-thinking Assessment Test – Stein et al., 2007), was chosen because of the disclosed scoring process, training availability and also the question format (requiring written responses and explanations, and testing multiple skills rather than testing skills individually).

- The student perception survey was adapted from the Student Assessment of their Learning Gains instrument (freely available). To align the questions to the CT performance measure, the skills directly measured in the test were also included in this survey.

- The observation tools were selected because they are freely available, easy to use, and the primary observation tool (Classroom Observation Protocol for Undergraduate STEM - Smith, Jones, Gilbert, and Wieman, 2013) came with a short training module which enabled the coding to be more consistent than other qualitative approaches, and therefore more comparable.

All tools were selected because of the thinking opportunities they could evaluate or inform. However, another part of the key criteria for choosing instruments included their availability and ease of use. I wanted to create a system that was accessible and meaningful for educators who may not have sufficient training or expertise in CT to make these decisions. To increase cohesiveness across the instruments, perception survey questions were aligned to skills directly testable by the CAT and were used to collect information from students and educators about the perceptions of the activities in the courses and the skills they believed they developed. Investigating the change in performance and perceptions enabled comparisons of the actual and perceived development of CT skills. This combination of the CAT and a modified SALG to inform classroom understanding about CT development is similar to the work of Styers, Van Zandt and Hayden (2018). However, the addition of the observation instruments, the Australian context and type of science courses examined distinguish my approach from theirs.

The other main issue to consider was the length of the study. A meta-analysis by Niu et al. (2013) revealed that single interventions of at least 12 weeks in length seemed to have the greatest effect on CT development. Other studies of CT development tend to be longitudinal or compare cross-sections, but this tends to give institution-level data, not student or classroom specific data. Given this research was interested in exploring the effects of different science classrooms, and the fact that there is not already a large Australian data set available to compare to, a semester-long pre-test post-test design was considered the best option to achieve these goals.

4.4.1 Instrument specifics

Critical-thinking Assessment Test - CT performance measure

Measurements of student CT skill development were made using a general CT test called the Critical-thinking Assessment Test (CAT). This instrument is a short answer essay-style test that uses real-life scenarios to assess various thinking skills (Stein, Haynes and Redding, 2006). The skills the CAT tests for include: evaluating information; interpreting information; creative thinking; problem-solving; communication (Stein et al., 2007). This instrument was developed at Tennessee Tech for the purpose of helping close the assessment loop for CT – giving educators understanding and experience in evaluating CT, as well as access to CT performance results for their students. Only one version of the test was available at the time of data collection, so the same test was used each time. This was not ideal, but the test creators report no “significant improvements…for control groups” who had been administered the test twice “over a semester or less” (Centre for

Assessment and Improvement on Learning, 2013, p.17). The CAT was administered in class at the start and end of the semester. Students were only given the opportunity to sit the CAT a maximum of twice.

In many ways the CAT was comparable to other standardised assessment instruments; however, the CAT's educator-involved scoring method and model for professional development set it apart from the other tests (Table 4.1). The tool generates an overall score for CT performance and sub-scores for each question (see Appendix F for a disclosable sample question and scoring explanation). It also provides a suggested theoretical guide for interpreting results, indicating which questions relate to a particular CT skill-set ("evaluate and interpret information; problem-solving; creative thinking; effective communication," Centre for Assessment and Improvement on Learning, 2013, p.35). Getting educators involved in the scoring process not only gives direct insight into student performance but has been found to build a stronger connection between the results and implications for practice (Stein and Haynes, 2011). Further, when educators from different disciplines collectively participate in scoring the tests, it generates wider discussion about CT (Lisic, 2015). In terms of more specific professional development, the CAT team train representatives from each institution to lead scoring workshops. They also offer training in writing questions which are similarly structured to the CAT. This is not intended to result in a situation where educators would be teaching to the test. Instead, it is expected that educators could use this training to embed discipline-specific CT practice within their courses, and then use the CAT to see if the skills are transferrable to more general contexts (Haynes et al., 2016). I undertook both training types, and incorporated practice questions as an intervention in one semester of the biology and society course to explore the effect on CT development (see Chapter 5. Section 5.3).

Table 4.1 Comparison of the tertiary level CT tests considered for use in the case studies.

Instrument	Question focus	Question format	Scoring	Professional development	Reference paper
California Critical Thinking Skills Test	General skills	Multiple choice	External processing	No	Facione et al. (1990c)
Collegiate Learning Assessment (CLA)	General skills	Short answer	External processing	No	Klein et al. (2007)
Critical thinking Assessment Test (CAT)	General skills (but science-themed questions)	Short answer	Completed by educators, verified by creators	Yes	Stein, Haynes and Redding (2006); Stein et al. (2007)
Cornell Critical Thinking Test Level Z	General skills	Multiple choice	External processing	No	Ennis, Millman and Tomko (1985)
Ennis-Weir Critical Thinking Essay Test	General skills	Essay	Completed by educators	Manual	Ennis and Weir (1985)
Watson-Glaser Critical Thinking Appraisal	General skills	Multiple choice	External processing	No	Watson and Glaser (1980)

Student perception survey

The student assessment of their learning gains (SALG) is an instrument with Likert-style questions used to gather "learning-focused feedback" from college students (SALG website, 2016). At the time of the student perception survey development, questions asked students about understanding, skills, attitudes, learning experiences, integration of learning, and grade point average. Inspired by the ideas in the SALG, a CT perception survey was developed to investigate another aspect of CT development. Using statements about skills reportedly measurable on the CAT, this self-developed survey asked students about skills the course let them practice and which skills they think improved. The final version of the survey and survey key can be found in Appendix D. Chapter 5 Section 5.2.2 and Appendix D also highlight modifications made between piloting this survey in Biology and Society Version 1 (post-test only) and its subsequent deployment in Chemistry and Biology and Society Version 2 (where additional items were added and both a pre-test and post-test version were completed by students).

Classroom observations

To close the feedback loop, educators need performance measures (Stein and Haynes, 2011) and insights into their practice. However summative evaluations often lack the diagnostic information needed to meaningfully transform instruction (Wiliam and Black, 1996; Wiliam and Black, 2018). Similarly, trying to understand CT development purely CT performance test results is not sufficient, because a score cannot explain why students performed the way they did. Additionally, because of the complex nature of learning, most tests are unable to help educators understand the specific aspects of their teaching which produced the effect or hindered the development of CT. To better inform interpretations of CT assessments and more deeply explore the opportunities for CT development within each case study, observations of classroom experiences were documented using a combination of published observational tools. These tools included one quantitative behavioural measure – the Classroom Observation Protocol for Undergraduate STEM (COPUS) (Smith et al., 2013), and two qualitative measures: the Electronic Quality of Inquiry Protocol (EQUIP) (Marshall et al., 2009); and Queensland School Reform Longitudinal Study protocol (QSRLS) (The School Reform Longitudinal Study Research Team, 2001).

Summaries about each of these tools can be in Table 4.2 (with copies of the instruments in Appendix E). All were developed to help educators explore their pedagogy and identify areas for improvement. This combination of tools enabled the observer to capture how time was spent; reflect on the nature of the presentation of knowledge, as well as reflect on the thinking demands offered in each course. In particular, the combination of the COPUS approach with the CAT data, made an ideal method for exploring if any particular pattern of teaching approaches led to stronger positive outcomes in CT development. For these reasons, it became the primary tool for data collection. However, a few questions from the EQUIP[11] and QSRLS tools were also added to capture some of the thinking demand in the courses (which the COPUS codes missed). Other studies have also modified the COPUS protocol to make it more suitable for purpose (Evenhouse et al., 2018). There will always be a gap between enacted curriculum and the experienced curriculum, however, this set of tools was intended to maximise the opportunity to identify how

[11] *Note:* The EQUIP instrument was initially trialled as the main data collection tool; however, a few pilot sessions using the EQUIP protocol revealed that it was quite hard to code quickly (I tried both the paper template and the IPad application). During piloting it also became evident that many of the EQUIP codes were not going to supply sufficient overview of classroom experiences, even though some of the aspects of the instrument were valuable to my study.

the actions of an educator or course can impact CT development. The other benefits of these tools are that they are all freely available; and the main tool used was developed for STEM (science, technology, engineering and mathematics) environments and has detailed coding instructions and a training protocol (see Appendix E), so educators can readily take this method and use it to examine their own classroom. Observations of the courses were made live, and also via video recordings of classes when available.

Table 4.2 Tools/instruments used for classroom observations. Copies of these instruments can be found in Appendix E.

	Tool	Description	Purpose for this study	Protocol modifications
Quantitative	*Classroom Observation Protocol for Undergraduate STEM (COPUS)*	The COPUS tool uses a simple coding system to capture the range and frequency of teaching tactics used within a course.	To collect data about how time was spent in class.	• Observations were made continuously throughout the semester, at each class time as opposed to the Smith et al. (2013) protocol where one or two random observations of the course would be made. • The full 24 code system was used when making observations; however, they were not all used in the final analysis. These changes are explained below in Section 5.4.2: Data processing. • The other amendment to the COPUS protocol was that all coding was done by one individual.
Qualitative	*Queensland School Reform Longitudinal Study (QSRLS) classroom observation protocol**	This tool uses descriptors with a scale of 1-5 to document evidence about the overall nature and style of the learning opportunities offered in the classroom.	To help build a profile about the construction and evaluation of knowledge in the courses.	• The dimensions included in the case studies were – Knowledge integration (p.3); Problematic knowledge (p.4); Problem-based curriculum (p.19) (The School Reform Longitudinal Study Research Team, 2001). • Coding per these QSRLS dimensions was added directly after COPUS coding, to capture the general treatment of knowledge in each class. • Due to the different types of learning experiences offered within each course, QSRLS data is presented per activity type (such as lecture, tutorial, workshop, practical).
	Electronic Quality of Inquiry Protocol (EQUIP)	The instrument focusses on documenting inquiry experiences in the classroom.	To provide a consistent way to document classroom experiences which align to CT and were not covered by the other tools.	• The tool was used in a reflective format, instead of live coding. While more subjective, the creators of the tool did note the tool can be validly used in this way (Marshall et al., 2009). • Only the discourse factors and the cognitive levels were used because the COPUS and QSRLS instruments covered other aspects.

* The QSRLS tool was originally designed for use in schools rather than in a university context; however, some of the dimensions are general enough to be relevant to other education settings.

4.4.2 Chemistry case study method

Experimental approach

The data collection approach for student perspectives and performance consisted of a pre-test post-test method in a quasi-experimental design. Data was collected from a single semester, from a subset of 78 of the 218 students studying the course. These particular students were enrolled in the Tuesday morning or Tuesday afternoon workshops, which were led by the two academics who developed the course. Students completed a standardised CT assessment and a perception survey at the start and end of the semester. The student perception survey was completed online through SurveyMonkey®. CT performance was assessed using the paper version of the Critical-thinking Assessment Test (CAT)[12].

Classroom observations were made throughout the semester (both live and through video recordings), capturing the learning opportunities between the pre-test and post-test performance. Observations commenced after the students had completed the first CAT test in week 2's workshop, and ended upon completion of the CAT in week 12's workshop. Workshop observations were made live, and lectures were coded from recordings of the live lectures. However, laboratory sessions were coded based on discussions with teaching staff, since the layout of the lab was not suitable to enable non-disruptive observations of interactions between teaching staff and students.

Data processing

<u>Participant data</u>

Research IDs were used to pair pre-test and post-test data within and across the tools. Participants who did not complete the pre-test and post-test of the CAT were excluded from all analyses. 69 matching CAT tests pairs were identified and extracted for scoring. This exceeds the paired-sample requirement suggested by the CAT developers (minimum of 15 pairs). Next, the paired perception survey responses were checked for completeness (how many question-sets and items were skipped). Responses with one or more skipped question-sets, in either the pre-test or post-test, were excluded from the perception survey analysis. This reduced the perception survey sample to 63. Responses with skipped items were treated as missing data and were excluded on a case-wise basis.

[12] Due to the use of the imperial measuring system in the USA, a glossary was provided with the CAT instrument to aid with question interpretation.

CAT scoring

The CAT tests would normally be scored by a team of 10-12 educators in one day. However, for this course, the CAT tests were scored by me and the two academics that ran the workshops. Scoring was completed over a number of shorter 2-3-hour sessions to accommodate for other commitments. The test creators were consulted before employing these variations, with the suggested strategy being to score related questions in chunks. An example of the scoring system is provided in Appendix F.

COPUS protocol modification[13]

While the full 24 COPUS coding system was used to collect the observation data, only 13 codes were used the final analysis. This involved merging 17 codes into 7, and excluding two codes. A complete summary of these code changes and justifications is displayed in Table 4.3.

[13] Modifications to protocols for other observation tools (QLSRS and EQUIP) were previously described in Table 4.2.

Table 4.3 Summary of COPUS code merges and exclusions for analyses

	COPUS description (Smith et al. 2013)	COPUS code	Modified code	Justification for modification
Student related codes	Student answering a question posed by the instructor with rest of class listening Making a prediction about the outcome of demo or experiment	AnQ Prd	AnQ	Both codes referred to an activity where students would be responding to an instructor prompt with other students listening.
	Individual thinking/problem solving. Test or quiz	Ind TQ	Ind	Descriptors for both codes specify individual thinking time.
	Listening to instructor/taking notes, etc.	L	L	
	Waiting (instructor late, working on fixing AV problems, instructor otherwise occupied, etc.) Other – explain in comments	W O	O	Both codes related to non-learning class time (such as being on a break or waiting for the instructor to set something up).
	Presentation by student(s)	SP	excluded	There were no student presentations within either course.
	Student asks question	SQ	SQ	
	Engaged in whole class discussion by offering explanations, opinion, judgment, etc.	WC	WC	
	Discuss clicker question in groups of 2 or more students Working in groups on worksheet activity Other assigned group activity, such as responding to instructor question	CG WG OG	WG	Counts for these three group-work codes were summed together to portray a single story about class time spent working in groups.
Instructor related codes	One on one extended discussion with one or a few individuals	1o1	1o1	
	Listening to and answering student questions with entire class listening	AnQ	AnQ	
	Showing or conducting a demo, experiment, simulation, video, or animation Presenting content, deriving mathematical results, presenting a problem solution, etc. Follow-up/feedback on clicker question or activity to entire class	D/V Lec Fup	Lec	Codes all represented the instructor presenting some form of content to the whole class.
	Moving through class guiding ongoing student work during active learning task	MG	MG	
	Administration (assign homework, return tests, etc.) Waiting when there is an opportunity for an instructor to be interacting with students Other – explain in comments	Adm W O	O	All codes related to non-learning class time (i.e. setting up an activity/computer, passing out handouts, or students on a break).
	Posing non-clicker question to students (non-rhetorical) Asking a clicker question	PQ CQ	PQ	Both codes refer to an activity where the instructor is posing a question to the class.
	Real time writing on board, doc. projector, etc.	RtW	excluded	Code occurred concurrently with 'Lec' and/or 'FUp' and therefore did not add new information to classroom profiles.

Descriptive analyses and statistical approaches

Due to having multiple data collection tools, each case study had multiple analyses (see Figure 4.1 for the analyses for the chemistry case study).

Approaches for observation data

To capture the potential CT development experiences offered to students the descriptive and qualitative analysis of the range of observation tools employed, observation data was considered in two ways. The quantitative COPUS data was used to generate profiles on how student and instructor time was spent. This was completed in Microsoft Excel (v.2010). Next qualitative summaries concerning how knowledge was treated (coded as per QSRLS protocol) were generated to build a further picture of the possible student learning experiences. These summaries were constructed per classroom experience: lecture, lectorial, workshop, and practical.

Summaries were also generated for the more qualitative analysis instruments. For example, a reflection on the discourse factors and cognitive levels (as per the EQUIP protocol) was undertaken, followed by examination of the treatment of knowledge. Both these instruments were scored using the protocols outlined in Appendix E.

These three classroom observation insights were used to help inform the CAT performance results by providing some indication of whether particular activities or questions seem to correlate to a greater change in CT performance. If there was a greater change, it was considered indicative that the approach promoted more CT development in students.

Approaches for CT performance and perception data

To focus more specifically on CT development, student perceptions and performance were explored. CT performance (as determined through CAT scores) was used to infer changes in CT ability. Student perception survey data was then used to investigate ideas about: how students perceive their CT abilities; if they thought the course helped develop them; and how they thought they use these skills on a daily basis (Appendix D). Paired-hypothesis testing was undertaken for performance and perception assessment tools to explore for differences in pre-test and post-test values within each tool type using GraphPad Prism (v.7). The particular hypothesis test employed depended on the question scale. For example, Paired T-tests were employed to investigate the significance of changes in overall CT level at the 95% confidence level for the overall CT performance. However, different tests (such as a two-sided chi-squared or Wilcoxon signed-rank test) were used to explore changes in individual CT skill performance as appropriate to the question scale (see Appendix H: Table H.1). Student perceptions of opportunities to practice CT skills during the semester were only on the post-test (Appendix D: Table D1). For analysis, the 5-point scale was reduced to 3 (regular = every class + weekly; irregular = a few times + once; and never). These categories were summarised into proportions and analysed in the combined analysis (described next).

Lastly, to explore patterns across the assessment tools (performance, skills perceptions and practice perception), independent sample-similarity matrices were generated by comparing the difference between pre-test and post-test survey and CAT test responses. Due to the presence of some negative values, score differences were transformed into a three-part coordinate. One variable indicated the magnitude of the score difference, and then the other variables used to indicate the direction in the form of presence/absence data (i.e. positive score: y/n, negative score: y/n). These independent sample-similarity matrices were then correlated using Spearman rank correlation (Clarke and Warwick, 2001) via the Mantle Relate test in PRIMER (v.6) with 9999 permutations (Figure 4.1). Further statistical analyses could not be undertaken due to differences in scales across the various tool types. For the multivariate analyses in PRIMER missing data was excluded on a case wise basis.

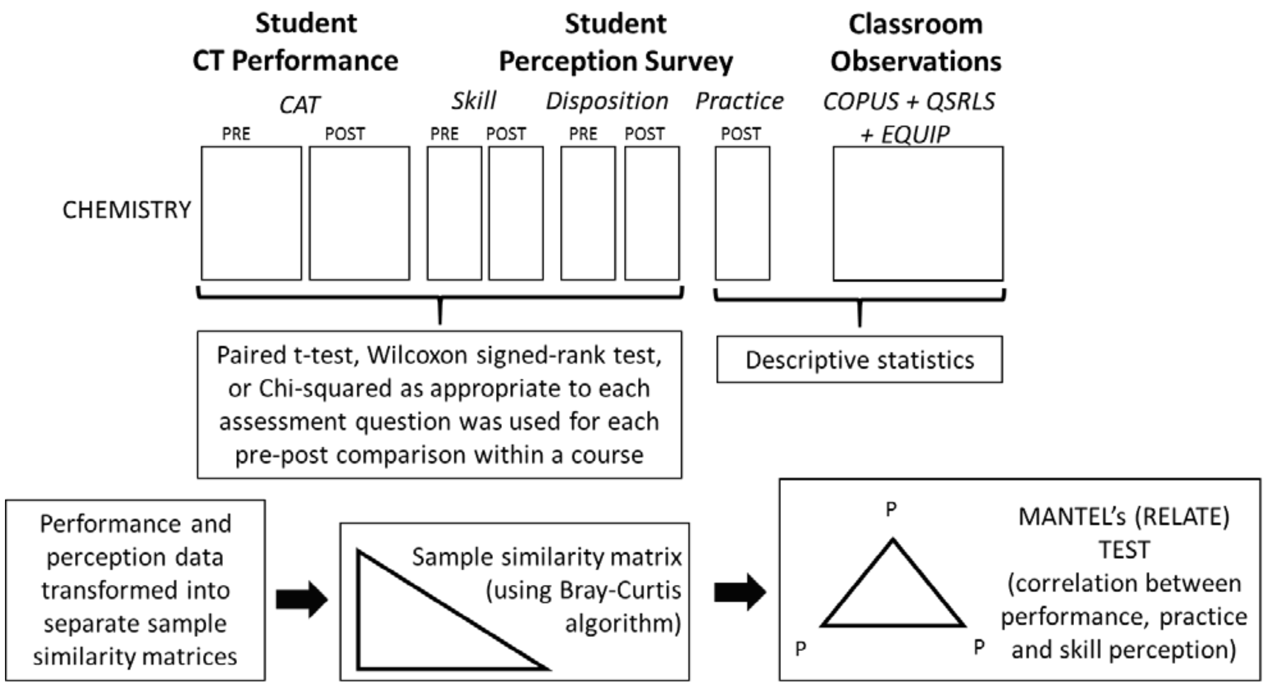


Figure 4.1 Summary of the statistical analyses undertaken within the chemistry course

4.5 Chemistry course structure and background information

This case study explores the effect of a discipline-specific science course (first year chemistry) on the development of CT skills. Students studying this course attend two 1-hour lectures a week, a fortnightly 3-hr workshop and a fortnightly 3-hour laboratory session. Figure 4.2 shows there was a fairly even split of course time across the three activity types. The lectures were delivered by two senior lecturers, with additional staff to support workshops and lab sessions. However, to minimise variability in teaching styles for the purpose of this study the workshop sessions observed in this this study were only those delivered by the senior lecturers. The lecturers actively engage with education research and employ teaching methods including group discussions and case studies in their teaching. The learning outcomes for this course include developing "problem solving strategies" and efficient "communication" (Koeper, 2015, p.2 – see Appendix G: Table G.1). The assessment has also been aligned to graduate qualities which include "graduates who… can apply their knowledge… and connect across boundaries" (Koeper, 2015, p.3-4 – see Appendix G: Table G.2). The course coordinator used an embedded CT pedagogy, but mostly anticipated CT would be an emergent property of the learning experiences. However, the course outcomes do include language associated with CT, including the terms 'analyse,' 'explain' and 'connect across boundaries.'

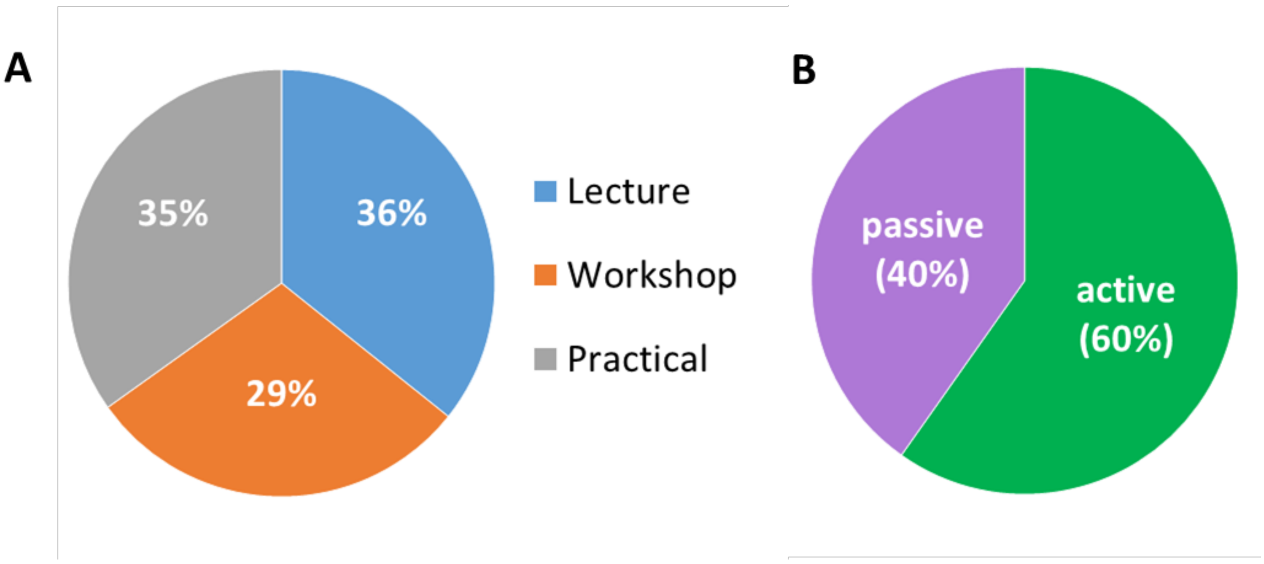


Figure 4.2 Summary of the portion of class time allocated across the various components of the chemistry course (A), as well as the proportions of active learning time in the course (B).

4.6 The effects of a first-year chemistry course on CT development

The results are summarised starting with specific trends emerging from explorations of classroom experiences, student perceptions and performance, before moving to a broader comparison across each of these datasets. Overall, the results demonstrate the sum of experiences in the first-year chemistry course was positively correlated with observed improvements to CT skill performance.

4.6.1 Summary of classroom experiences

Observation tools findings: COPUS

Summaries of the COPUS results revealed that listening was the most frequently observed student behaviour in the chemistry course, but collectively more time was spent in an active learning-mode than a passive one (Figure 4.3).

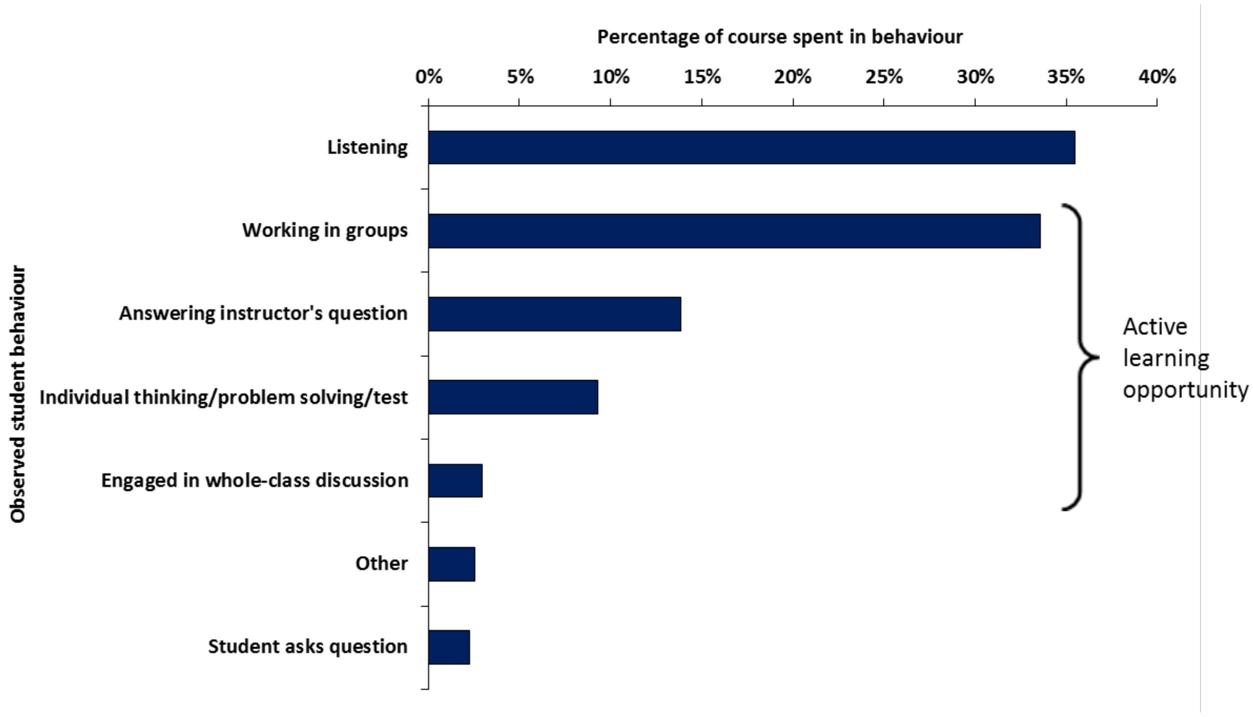


Figure 4.3 Summary of how student time was spent in the chemistry course.

In general, the patterns of how time was spent were reflective of what educators would typically expect of a lecture, workshop/tutorial and laboratory environment (Figure 4.4 A and B). That is to say, the workshops included the broadest range of student actions (including making predictions, taking quizzes and working in groups); lectures involved a lot of instructing and listening; and practicals involved the least amount of instructor guidance and the most amount of group work. Instructor time was used effectively to engage students with course content, with the total sum of experiences comprising of >5% of 'other' non-content related activities (Figure 4.4B). In fact, the majority of this 'other' time represented the short break given to students during each three-hour workshop timeslot. A final aspect to note is that even though the lectures and workshops were delivered by two academics, there was little difference in the way these two instructors presented their content.

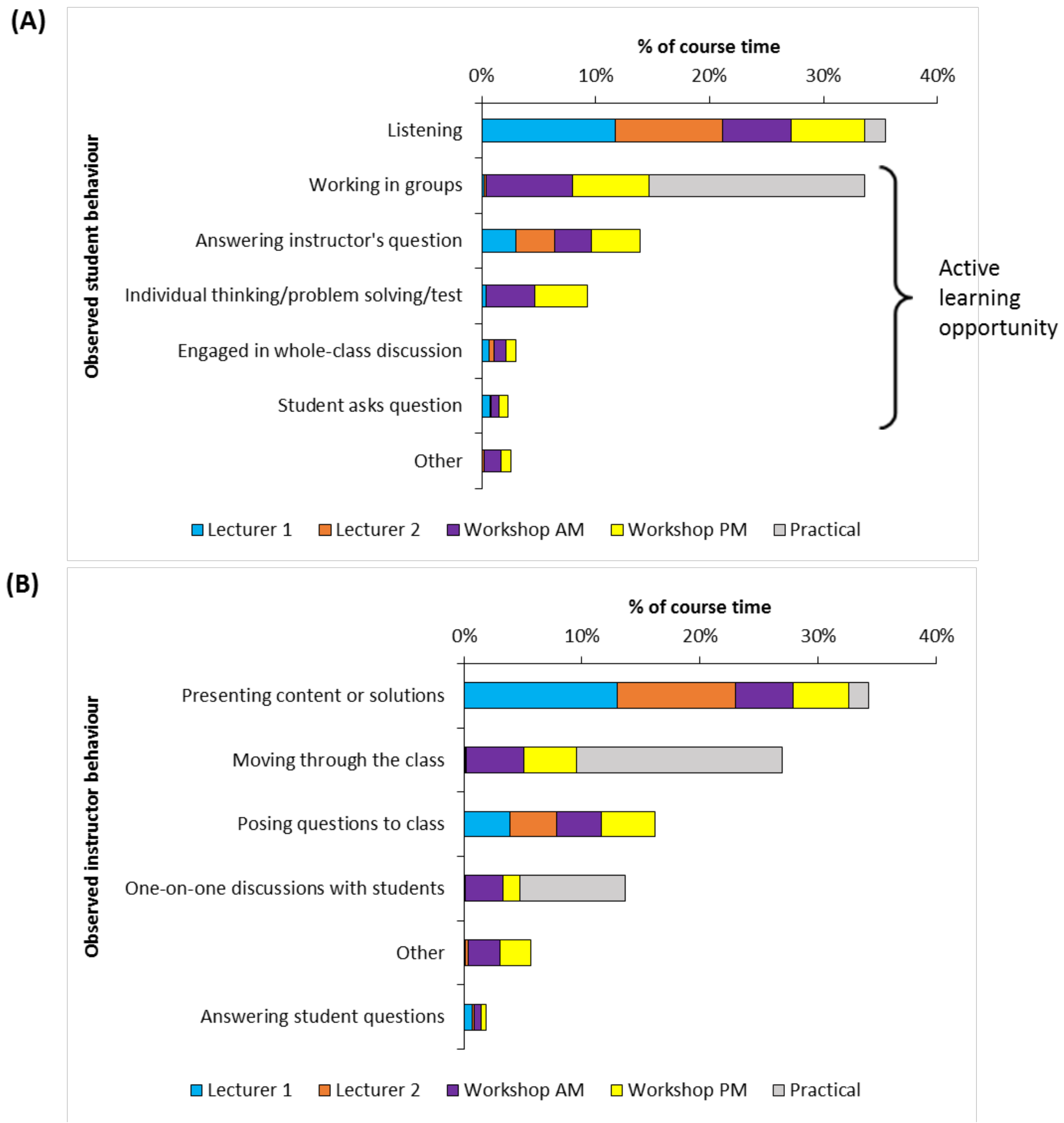


Figure 4.4 A-B Summary of student and instructor behaviours in each activity type in the chemistry course. Where (A) shows how student time was spent in each course activity and (B) shows the breakdown of educator time in each course activity.

Observation tools findings: QSRLS

Next, to further explore the contributions of the lecture, workshop and practical components of the course on CT development, aspects of the QSRLS were incorporated to categorise how knowledge was treated in the classroom. A score between 1 and 5 was possible for each dimension, with the protocol dictating that all expectations contained within each descriptor must be met to achieve that descriptor's score (see Appendix E).

In terms of the QSRLS dimension 'knowledge integration', knowledge was assessed as a score of 2 – "mostly restricted to that of a specific subject area" (The School Reform Longitudinal Study Research Team, 2001, p.3). The course rarely discussed other disciplines. Occasionally, relevant physics principles were mentioned, however, this was generally to help student taking physics to understand any distinctions in the required knowledge for the chemistry course.

For the dimension 'problematic knowledge,' the score assigned varied depending on the activity type. For this course, lectures were given a score of 2. Even though the lecturers would present facts or findings and then ask why it was so (a possible score of 3); it was only scored as 2 because the given answers were expected to be "reducible to a given body of facts" rather than considering the possibilities of the social construction of knowledge. Given this is a core foundational science course for the student's degrees, it is not overly surprising that the lectures were handled in this fashion. However, the range of activities within the fortnightly workshops meant these experiences were more closely aligned to the descriptors for score 3, where "multiple interpretations" of information were discussed around a core set of chemistry principles (The School Reform Longitudinal Study Research Team, 2001, p.4). Practical experiences consisted largely of students following a cookbook type laboratory lab manual with guidance available from a teaching assistant. Depending on the student this meant that this activity operated at a score of level 1 or 2 in the 'problematic knowledge' space, as students were generally not challenged to go beyond the observed "body of facts" (The School Reform Longitudinal Study Research Team, 2001, p.4).

The final aspect used from the QSRLS protocol was 'problem-based curriculum.' Workshops supplied the greatest number of opportunities for students to encounter creative problem-solving tasks. Part of the regular fortnightly assessments required students to solve an unfamiliar real-world context problem that was based on the content they had previously covered in class. This required both "knowledge construction and creativity" as students explored the scenario and sought out an answer to a “small” problem (The School Reform Longitudinal Study Research Team, 2001, p.19). However, students were also given many opportunities to solve more content-based problems in lectures and in practical sessions. Most problems that were set in lectures and workshops needed just a few minutes to solve, with larger problems never taking more than 25% of total class time to solve. All but one of the laboratory sessions were structured in a cook-book style, so even though students spent the majority of class time running experiments and doing calculations, the overall intellectual demand was similar to the short problem-based activities undertaken in the workshop. The majority of the problems presented in this course scored a 2 on this dimension. However, some of the workshop tasks could be scored at 3, and the only non-cook book laboratory session (where students designed their own experiment to identify an unknown) warrants a score of 4.

Observation tools findings: EQUIP

The EQUIP tool was not used to make detailed classroom observations, but reflections of the classroom experiences against the discourse factors were made (as per suggestion by Marshall et al., 2009, p.49) and these revealed that the course generally functioned at the developing inquiry level (see Table 4.4). Most questions were at a level that tested student understanding, with one correct answer, and rarely challenged them into the application level. However, students were expected to be able to justify their answers with the relevant evidence (essentially drawing on content that had previously been covered in class). The fortnightly workshop was where most questions were asked. But the capacity to record question phrasing or code accurately in this setting was limited due to the class only being available for live coding. However, it was noted that there were regular expectations in the workshops to answer questions (or follow-up questions) to "justify reasoning," thus shifting some of the classroom experiences up to the proficient inquiry level (Marshall et al., 2009, p.49). The other aspect of the EQUIP tool which was considered was the cognitive demand on students. In this course 5 of the 6 cognitive demand levels were witnessed by the researcher; with the majority of time being spent in the "receipt of knowledge" and "lower order" cognitive stages.

Table 4.4 Summary of the EQUIP results for chemistry

EQUIP component		Level	Levels displayed		
			Lecture	Workshop	Practical
Cognitive Level		Other	✓ *	✓ *	
		Receipt of Knowledge	✓	✓	
		Lower Order	✓	✓	✓
		Apply		✓	✓
		Analyze/evaluate			✓
		Create or transfer			✓ *
Discourse factors ~	Questioning Level	Preinquiry			
		Developing Inquiry	✓		✓
		Proficient Inquiry		✓	
		Exemplary Inquiry			
	Complexity of Questions	Preinquiry			✓
		Developing Inquiry	✓		
		Proficient Inquiry		✓	
		Exemplary Inquiry			
	Classroom Interactions	Preinquiry			
		Developing Inquiry			✓
		Proficient Inquiry	✓	✓	
		Exemplary Inquiry			

*observed but not regularly
~most common level achieved displayed

Student perceptions of classroom experiences

Students were asked to report on their perceptions of the learning opportunities in the chemistry course. At least 79% of students indicated that all 23 learning opportunities the survey included were practiced at least once during the semester (Table 4.4A and B). In fact, there were only four skills (all of which related to CAT measurable skills) that >50% students reported practising less than weekly (Table 4.4A Q2, Q5, Q6, Q8). With the exception of the responses to Q8 'determine whether an inference in an advertisement is supported by information' (description of skill measured by the CAT, Center for Assessment and Improvement of Learning, 2013) perception results were generally consistent with classroom observation results. This item (Q8) had the most ambiguous of the perception results, where 35.5% reported practising this at least weekly, 43.5% thought they only practiced it a few times at most, and 21% thought they never got to practice this skill in this course. This split is likely because one component of a workshop activity was testing a marketing claim (even though this was achieved by calculating the chemical content, rather than from the information in the claim itself). It speaks to the need to be clear about the terminology used in classrooms, so students can distinguish between the discipline-specific and general skills they are given the opportunity to learn. Another example demonstrating the importance of terminology is the result for Q23 (Table 4.4B). This statement referred to how often they got to practice interpreting and drawing conclusions from scientific data. A higher proportion of students thought they got to practice this skill weekly, than the more generic description of this same skill as assessed by the CAT (Table 4.5A and B). This suggests they may not have recognised that the skills described in Q23 were actually a similar skillset to Q4/Q7. But it also calls to question the amount of attention (and perhaps effort) that students put into completing the survey.

Table 4.5 (A) Summary of student perceptions of the CT related skills they thought they got to practice in the chemistry course.

Survey descriptor: CAT measurable skills		Perceptions of learning opportunities		
		practised regularly	practised irregularly	never practised
Q1	summarize a pattern of information without making inappropriate inferences.	59.7%	40.3%	0.0%
Q2	evaluate how strongly correlational-type data supports a hypothesis.	47.6%	47.6%	4.8%
Q3	provide alternative explanations for observations.	50.8%	49.2%	0.0%
Q4 / Q7*	identify additional information needed to evaluate a hypothesis or particular explanation of an observation.	55.6%	44.4%	0.0%
Q5	evaluate whether spurious relationships strongly support a claim.	33.9%	59.7%	6.5%
Q6	provide alternative explanations for spurious relationships.	25.4%	66.7%	7.9%
Q8	determine whether an inference in an advertisement is supported by information.	35.5%	43.5%	21.0%
Q9	provide relevant alternative interpretations of information.	53.2%	45.2%	1.6%
Q10	separate relevant from irrelevant information when solving a real-world problem.	66.7%	31.7%	1.6%
Q11	analyse and integrate information from separate sources to solve a real-world problem.	58.1%	38.7%	3.2%
Q12	use basic mathematical skills to help solve a real-world problem.	88.9%	11.1%	0.0%
Q13	identify suitable solutions for a real-world problem using relevant information.	65.1%	33.3%	1.6%
Q14	identify and explain the best solution for a real-world problem using relevant information.	59.7%	37.1%	3.2%
Q15	explain how changes in a real-world problem situation might affect the solution.	58.7%	38.1%	3.2%

*these two CAT questions have the same skill description in the CAT manual (see Centre for Assessment and Learning, 2013 p. 23).

Table 4.5 (B) Summary of student perceptions of the other skills they thought they got to practice in the chemistry course.

Other course skill description		Perceptions of learning opportunities		
		practised regularly	practised irregularly	never practised
Q16	explain the methods of science.	65.1%	34.9%	0.0%
Q17	consider why scientific knowledge is testable.	63.5%	33.3%	3.2%
Q18	explain why scientific knowledge is testable by further inquiry.	64.5%	33.9%	1.6%
Q19	explain the role of science in society.	57.1%	42.9%	0.0%
Q20	explain the relevance of science in society.	68.3%	31.7%	0.0%
Q21	design and plan an investigation	41.3%	55.6%	3.2%
Q22	collect and accurately record scientific data.	73.0%	23.8%	3.2%
Q23	interpret and draw conclusions from scientific data.	79.4%	20.6%	0.0%

Student perceptions of their CT and general science abilities

Students were asked to reflect on their CT abilities, as well as some more general science skills, at the start and end of semester. These questions aligned to the statements about the skills the course gave them opportunity to practice, as well as the skills the CAT could measure. Most students agreed with all of the statements in the pre-test, indicating they thought they already had some proficiency at these skills (Table 4.6 and Table 4.7). Only one of the 23 statements, Q5, was found to be statistically significantly different at the 0.05 level (Table 4.6). For this question, which referred to evaluating whether spurious relationships support a claim, there was a positive shift in student's assessment of their ability between the start and end of semester (Table 4.6). This skill was measurable on the CAT test and is discussed later, as student CT performance was also found to significantly improve at this question.

Table 4.6 Summary of the perception survey results for student judgments about the CAT measurable skills for the chemistry course. Significance explored using a Wilcoxon signed-rank test statistic (*W*), with two-sided p-value. Alpha was set to 0.05, where bolded p-values indicate where a significant difference lies.

	Q1		Q2		Q3		Q4		**Q5**	
	pre	post	pre	post	pre	post	pre	post	pre	post
Maximum	5	5	5	5	5	5	5	5	5	5
Median	4	4	4	4	4	4	4	4	4	4
Mean	4.00	4.05	4.00	4.02	4.14	4.07	4.14	4.23	3.49	3.85
Statistical test statistic	30		16		-45		45		**427**	
Statistical test: p-value§	0.069		0.898		0.729		0.421		**0.004**	

	Q6		Q7		Q8		Q9		Q10	
	pre	post	pre	post	pre	post	pre	post	pre	post
Maximum	5	5	5	5	5	5	5	5	5	5
Median	4	4	4	4	4	4	4	4	4	4
Mean	3.68	3.84	4.14	4.23	3.98	3.94	4.02	4.08	4.22	4.19
Statistical test statistic	137		45		-62		67		-59	
Statistical test: p-value§	0.189		0.421		0.672		0.512		0.476	

	Q11		Q12		Q13		Q14		Q15	
	pre	post	pre	post	pre	post	pre	post	pre	post
Maximum	5	5	5	5	5	5	5	5	5	5
Median	4	4	4	5	4	4	4	4	4	4
Mean	4.18	4.15	4.43	4.57	4.00	4.20	4.16	4.18	4.27	4.18
Statistical test statistic	-19		91		15		179		-87	
Statistical test: p-value§	0.873		0.202		>0.999		0.119		0.378	

Table 4.7 Summary of the perception survey results for student judgments about the general science abilities for the chemistry course. Significance explored using a Wilcoxon signed-rank test statistic (*W*), with two-sided p-value. Alpha was set to 0.05, where bolded p-values indicate where a significant difference lies.

	Q16		Q17		Q18		Q19	
	pre	post	pre	post	pre	post	pre	post
Maximum	5	5	5	5	5	5	5	5
Median	4	4	4	4	4	4	4	4
Mean	4.14	4.03	4.02	4.08	4.16	4.15	4.30	4.24
Statistical test statistic	-91		71		-19		-66	
Statistical test: p-value§	0.230		0.591		>0.999		0.639	

	Q20		Q21		Q22		Q23	
	pre	post	pre	post	pre	post	pre	post
Maximum	5	5	5	5	5	5	5	5
Median	4	4	4	4	4	4	4	4
Mean	4.43	4.29	4.16	4.12	4.35	4.39	4.18	4.34
Statistical test statistic	-136		-19		11		85	
Statistical test: p-value§	0.097		0.868		>0.999		0.147	

Student perceptions of their disposition to transfer their knowledge and skills

As part of the perception survey, students were also asked to reflect on four statements concerning how they transfer and apply their knowledge and thinking skills from one context to another. Irrespective of the point in the semester, most students agreed with these statements (Table 4.8). The extent of this agreement remained fairly consistent, with the exception being a 10% reduction in support for statement 1. There were no statistically significant shifts in student dispositions. There were a handful of students who did experience a positive shift in dispositions across the semester (Figure 4.5). However, what was more interesting in relation to CT development was that fewer students disagreed with the statements at the post-test, with a 5-15% shift toward a neutral stance (Table 4.8). This small positive shift for this subset of students is likely indicative of an increase in the degree to which they recognise they engage in CT processes.

Table 4.8 Summary of the disposition-related perspectives for the chemistry course.

	Statement	agree		disagree		unsure	
		pre	post	pre	post	pre	post
Q24	Presently I am in the habit of connecting key ideas I learn in my classes with other knowledge.	93%	83%	7%	2%	0%	15%
Q25	Presently I am in the habit of applying what I learn in classes to other situations.	84%	82%	15%	2%	2%	16%
Q26	Presently I am in the habit of using systematic reasoning in my approach to problems.	79%	77%	15%	0%	5%	21%
Q27	Presently I am in the habit of using a critical approach to analysing data and arguments in my daily life.	80%	79%	15%	3%	5%	18%

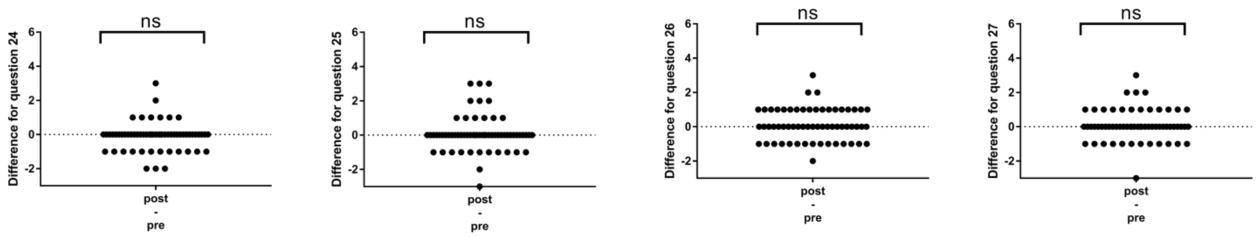


Figure 4.5 Summary of the chemistry student response differences for the disposition questions on the student perception survey.

4.6.2 Summary of CT performance results

When comparing the CT performance results for the chemistry cohort (n=69), significant differences were found between three individual questions (Q5, Q10 and Q11), as well as for the overall change in performance across the semester (Table 4.9). These individual questions assessed the skill "evaluating and interpreting information," as mapped by the CAT creators (Center for Assessment and Improvement of Learning, 2013, p.35), with Q10 and Q11 also assessing student "problem solving" skills. The test creators have also specified that Q11 assesses "effective communication" (Center for Assessment and Improvement of Learning, 2013, p.35). However, by nature of the question design, all short answer questions assess students written communication abilities (even if only some questions awarded extra points for the effective communication of an idea).

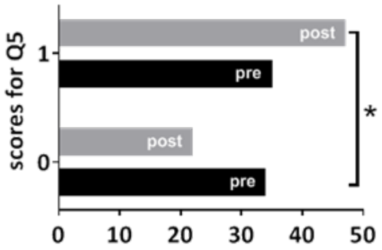
The performance differences on the CAT are indicative that the chemistry course helps to improve students CT performance. But is the finding of statistically significant differences for three questions and the overall test enough to support that this course has a positive effect on CT development? The study reveals that 63.5% of the student's overall performance change improved their performance score by more than 0.5, which was determined to be higher than what is typically expected from studying a semester of college[14]. In fact, the magnitude of the mean performance change observed for this single semester of first-year chemistry was around 2 marks (11%), which is just under 50% of the score increase the CAT creators' report is generally expected from a four-year degree (Harris et al., 2014). Despite the Australian students starting at a higher CT skill baseline, CT performance at the end of the chemistry course was significantly higher

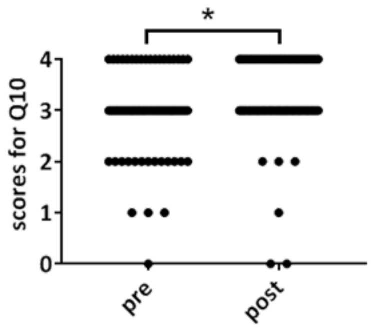
[14] Refer back to Section 4.3; see also Harris et al., 2014 for an explanation; refer to next section for a comparison between the Australian first-year chemistry cohort and US freshman norm scores.

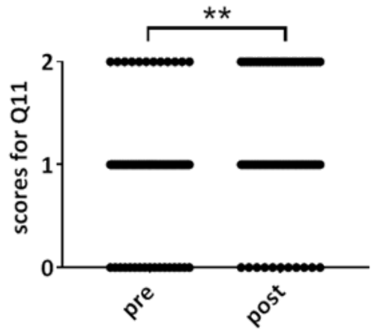
than at the start of semester. This improvement was also above what the general tertiary experience is thought to impart. The responses provided by students in post-test were also shorter than in the pre-test. Additionally, the minimum and maximum scores achieved in the post-test were slightly lower than in the pre-test. Both these factors are indicative of the students putting less effort into completing the post-test, because the parts of the CAT scoring system rewards longer answers and explanations (see Appendix F for further explanation of scoring).

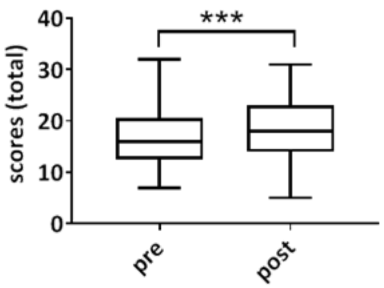
Further opportunities for improvement were evident in through exploration of the mean scores. This revealed that less than 50% of the possible points had been awarded on 7 of the 15 questions in the post-test (see Q3, Q4, Q6, Q7, Q9, Q14, and Q15, Appendix H: Figure H.1 and Table H.1). Performance on Q6 (concerning 'alternative explanations for relationships'), and Q7 (concerning 'additional information needed to evaluate a hypothesis'), decreased, though not significantly. Performance was lowest on Q4 (mean post-test score = 1.07; median =1), which like Q7, also concerned "identifying information needed to evaluate a hypothesis or particular explanation of an observation" (Center for Assessment and Improvement of Learning, 2013, p.24). Despite the fact that there was an overall significant increase in performance, there are still multiple opportunities for further improvement in student CT skills.

Table 4.9 Summary of the chemistry cohort's significant pre-test and post-test CAT results (n=69). Due to variations in the scales across questions, different statistical tests were applied as appropriate. These are indicated through the symbols near the question number and are explained below in the table notes. In all cases, alpha was set to 0.05, where bolded p-values indicate where a significant difference lies and the asterisk denote the level of significance, * <0.05, ** <0.01, * <0.001. A complete summary of the CAT results can be found in Appendix H: Table H.1 and Figure H.1.**









	Q5□		Q10^		Q11^		total#	
	pre	post	pre	post	pre	post	pre	post
Maximum	1	1	4	4	2	2	32	31
Median	1	1	3	3	1	1	16	18
Mean	0.51	0.68	2.94	3.23	0.93	1.22	16.49	18.32
Sum	35	47	203	223	64	84	1138	1264
Difference	12		20		20		126	
Statistical test statistic	4.328		336		316		3.469	
Statistical test: p-value	**0.037**		**0.019**		**0.003**		**0.0009**	

□ Chi-squared test statistic (chi-squared), with two-sided p-value. A contingency test comparing the proportions of 0 and 1 scores between pre-test and post-test.
^ Wilcoxon signed rank test statistic (W), with two-sided p-value. A non-parametric version of the paired t-test, to assess whether matched population mean ranks differ between pre-test and post-test.
# paired t-test statistic (t), with two-sided p-value. A parametric test after confirmation of normality (using the Shapiro-Wilk and D'Agostino & Pearson normality tests) to assess whether population means differ in the pre-test and post-test.

Comparison to American students

In general, student performance on both the pre-test and post-test was higher than expected based on the CAT creators' freshman norm data set (Figure 4.6). According to the report generated by the CAT creators (See Appendix H: Table H.2 and H.3) the Australian chemistry students and the freshman norm overall performance on the CAT were highly significantly different (pre-test and post-test p-values <0.001). As shown in Figure 4.7, there were only four questions where the Australian students performed lower than their freshman equivalents (Q1, Q4, Q5, Q7), and by the post-test there were only two questions where the American students still scored higher (Q1 and Q7). However, performance was only determined to be significantly lower (p-value <0.05) for Q1 on the pre-test (Appendix H: Table H.2), and Q4 on the post-test (Appendix H: Table H.2).

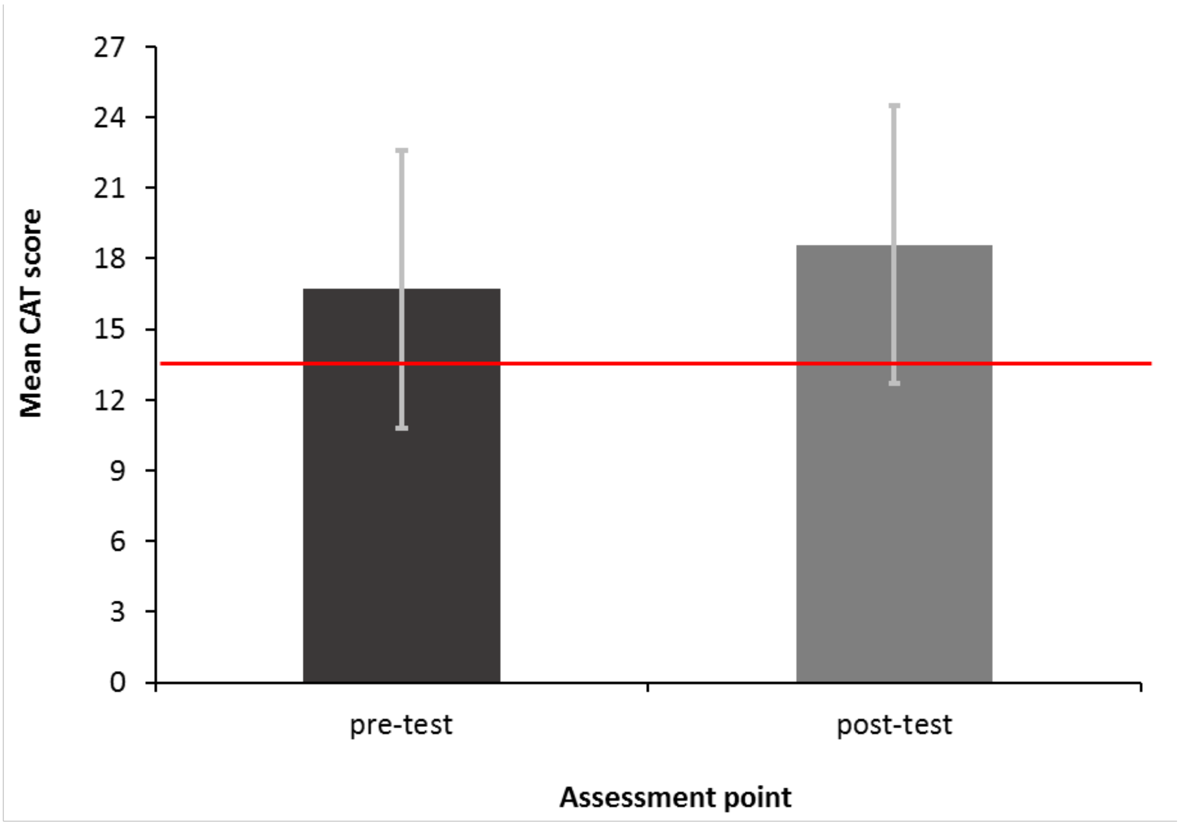


Figure 4.6 Mean CAT performance results at the start and end of the chemistry course. The maximum possible score is 38. The red line indicates the typical score of American Freshmen. The error bars represent +/- 1 SD.

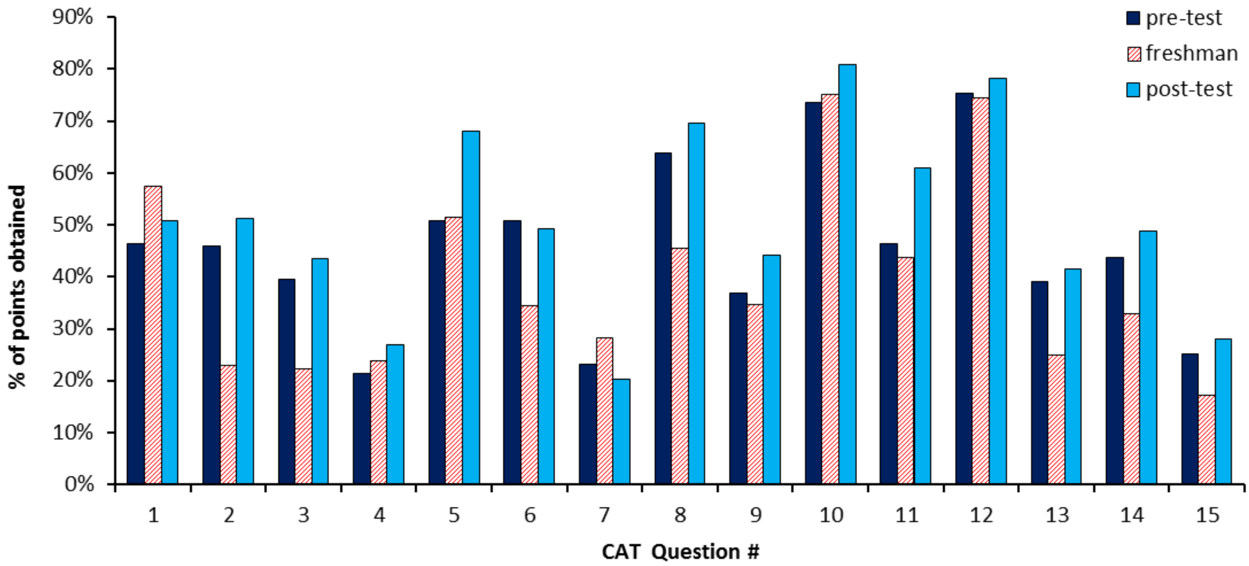


Figure 4.7 Summary of the mean % points obtained per CAT question for the chemistry cohort at the pre-test and post-test compared to their American freshman peers.

4.6.3 Comparisons of student perception of skills, practice and CT performance

The CAT performance results revealed that students improved on three CT skills (Q5, Q10, Q11, refer back to Table 4.9). However, when exploring the perception survey findings with the CT performance results, only a small subset of students (33.9%) accurately judged the improvement of their abilities for one of these skills (see Table 4.10: Q5). Overall, there were no correlations between the CAT skill perception, CAT skill practice and CAT performance responses - irrespective of looking at the data collectively, or by further exploring the trends for questions which had been determined to be significantly different. There were also no correlations between CAT performance and course grade, irrespective of pre-test performance or overall change on the CAT.

Table 4.10 Summary of the average (mode) changes in chemistry student's perception and performance across the semester for survey items where there was a significant change in CAT performance (n=69).[15] Students grouped by how frequently they thought they got to practice each skill, where the % value indicates the proportion of students who perceived that extent of practice. The symbols (↑ = ↓) display the average change in perception or performance for that grouping, where "↑" represents an increase, "=" no change and "↓" indicates a decrease/reduction. A complete summary of the results can be found in Appendix H: Table H.4.

		Practised regularly			Practised irregularly			Never practised		
Item	CT skill description (based on CAT measurable skills)	%	Change in perception of ability	Change in CAT performance	%	Change in perception of ability	Change in CAT performance	%	Change in perception of ability	Change in CAT performance
Q5^	evaluate whether spurious relationships strongly support a claim.	33.9	↑	↑	59.7	=	↑	6.5	↓	=
Q10	separate relevant from irrelevant information when solving a real-world problem.	66.7	=	↑	31.7	=	↑	1.6	↑	=
Q11	analyse and integrate information from separate sources to solve a real-world problem.	58.1	=	↑	38.7	=	↑	3.2	~	↑

^ This was the only item where both practice and perception stastisticall significantly imprroved at the post-test. ~BIMODAL- equal number of no change and increase

[15] Not all students who completed the CAT pre-test and post-test (n=69) completed the pre-test and post-test perceptions survey; however, there were a minimum of 63 responses per item.

4.7 Discussion

This chapter is a component of a set of studies designed to expand research about CT development in Australian tertiary science classrooms (part of aim 2), and trialling a multi-instrument approach that could serve as a general model for CT evaluation (part of aim 3). This particular chapter explores the effect of studying chemistry, a course whose embedded approach to CT development was representative of the key trends that emerged from my previous study (Chapter 3). Thus, the focus of this study was to investigate whether engagement with embedded CT opportunities through scientific methods leads to positive changes in CT development.

My studies are unique in their approach, each employing a trio of measures, comprising of 4 main data types. This approach was important for increasing the clarity and usability of the results, as one- or two-dimensional assessment regimes are reportedly not addressing educator's questions about CT development in their classrooms (Benjamin et al., 2016). The assessment tools were carefully selected so they could be used by educators who have not had extensive training in CT development, as a lack of professional development has also been identified as a key issue in the literature (Lauer, 2005; Black, 2009; Reynolds, 2016).

Whilst commentary on the American education system is often about over-assessing their students using standardised assessments (Heilig, Brewer, and Pedraza, 2018), the Australian system could benefit from strategically employing more formal evaluations, especially at a senior secondary and tertiary level where little to no formal explicit generic CT skills assessment takes place. Our schooling continues to drop in the world rankings (OECD, 2018; New Jersey Minority Educational Development, 2019). To rectify this, it is imperative for Australian educators to assess the extent of graduate attribute development (including CT development) in their classrooms. They need to know 1) what is working, and 2) what specifically needs to change so they can help Australian graduates keep pace with other first world nations.

4.7.1 Effect of first year chemistry on CT development

First year chemistry is an important keystone course in many Australian science students' degrees. The structure of Australian university degrees promotes an emphasis on disciplinary content. Unlike their US counterparts, Australian students start training in their chosen specialisation at first year. This means there is less opportunity to equip students with domain-general skills. Yet now more than ever, students need generic skills like CT to enable a career and make effective life choices (Moffit et al., 2011; Grieco, 2016; Bezanilla et al., 2019; Pearl et al., 2019). As a required first year course for the majority of science students, chemistry is well positioned to equip students with broader study and career skills, in addition to the foundational chemistry content knowledge needed for their degree.

Even though this chemistry course embedded CT among other tasks (i.e. what the literature suggest is the least effective way to develop CT), the course coordinator thought that the course would develop students CT skills. This belief was consistent with findings from Chapter 3, and themes in the literature around "hopeful pedagogy" (Nicholas and Raider-Roth, 2016, p.1; see also Paul et al., 1997b; Roades, Ricketts and Friedel, 2008). This belief also follows on from the idea that engagement with scientific methods leads to CT development (Lederman, 2008; Holmes, Wieman and Bonn, 2015). Indeed, this course did have a positive effect on CT performance development. In particular, students CT performance significantly improved on three questions, as well as on the overall test. Further, the extent of this performance change on the overall test was greater than 0.5-point score increase suggested by the CAT (assumption based on a linear change of 4 points for a four-year degree), suggesting that this course was able to improve students CT skills. Shifts in students CT abilities resulting from chemistry in Australia have not been evaluated before, thus this research adds to the body of knowledge about generic capability development in chemistry classrooms, specifically adding evidence that scientific approaches help foster students CT skills.

In line with wanting a more complete view of CT development, to determine success in CT development this study also factored in the affective aspects of CT - such as student's awareness of their abilities (perception shift) and their disposition to think critically. Whilst there were no correlations between the student's perceptions and performance, there was some growth evident in student perceptions of their CT skills. In particular, there was a significant change in student's belief about their ability to evaluate contentious claims, as well as in their performance in this skill. However, their perceptions about most of their abilities did not change across the semester. There were no significant shifts in students CT dispositions. Student responses to the dispositional items indicated a high level of agreement even at the start of semester, which could account for the lack of change. However, another reason why dispositional shifts may not have been seen is because dispositional changes may need longer to occur (Thompson, 2009). I think this is particularly true for embedded environments where non-disciplinary related reflection and metacognition may not be emphasised, encouraged or valued. However, the lack of dispositional shifts could also be attributed to the fact that a perception survey was used to evaluate this rather than a more formal assessment instrument such as the California Critical Thinking Disposition Inventory. Students are generally poor at self-evaluation of CT abilities (Harris, 2015; Zlatkin-Troitschanskaia, Shavelson and Kuhn 2015; Hyytinen, Toom and Postareff, 2018), so this result is not overly surprising. But it does mean there is still work to be done in this affective space. To become better critical thinkers, students and educators need to be able to recognise the strengths and weaknesses of their CT abilities so they can continue to develop them (Mason, Ariasi, and Boldrin 2011).

Lastly, there were also no correlations between CAT performance and course grade, irrespective of pre-test performance or overall change on the CAT. This was interesting because, previous studies have indicated that students with higher CT skills tend to perform better in chemistry (Fredrette, 2018). However, Fredrette's study was conducted using a different CT assessment (the Watson-Glaser Critical Thinking Appraisal), on a general chemistry class at an American college. Differences pertaining to the cohort, the education setting, as well as the CT test may account for the reason the same trend was not observed in my data. This course was able to improve CT performance, and to a small-degree, perceptions, suggesting that the collective sum of experiences in this chemistry course were enough to meaningfully improve students CT skills. However, a limitation of this study is that students were taking other courses at the same time. Note that further critique of my methodology is provided in Chapter 6 rather than here, to facilitate a more comprehensive review of case-study method and its ability to inform practice.

4.7.2 Practice implications

A specific objective of this research was to increase understanding about CT development in science classrooms in Australia. Having determined that this course did improve CT, it was time to examine what could have led to these improvements. Was it engagement with the scientific method? Or something else? The chemistry course was fairly traditional in structure, in that it consisted of lecture, practical and tutorial-like (workshop) components. The workshop environment was observed to offer a greater range of individual and group-based problem-solving experiences, and provided the most opportunity for engagement with CT processes compared to other course components. However, on the whole the course embedded CT skills within other learning tasks and did not seek to explain or emphasise CT in any explicit way. Although embedded teaching approaches reportedly have the smallest effect on CT development (Abrami et al., 2008, 2015; Tiruneh, Gu, De Cock, and Elen, 2018), this significant CT performance change is not necessarily surprising. The reasoning and mental procedures used for thinking scientifically are believed to be a subset of those used for non-scientific reasoning (Klahr, 2000). These mental procedures are also thought to help individuals to recognise the conditions of when to stop and start applying their reasoning skills (see Willingham, 2007). In addition, the instructors used a variety of instructional techniques to engage students in course content, as well as creating an active learning environment.

Active learning environments (ALE) are characterised by students engaging in tasks that require "to do more than just listen" and recall, they need to involve "analysis, synthesis and evaluation" (Bonwell and Eison, 1991, p.iii). ALE are long associated with developing thinking skills and influencing student's motivation and attitudes towards learning (Bonwell and Eison, 1991; Kim et al., 2013; Lumpkin, Achen and Dodd, 2015; Bezanilla et al., 2019), as well as improving conceptual understanding (Minner, Levy, and Century, 2009). Whilst some studies show that students prefer instructors to lecture 60% of a course (Achen and Lumpkin, 2015), lecture-based courses are more likely to result in failing an exam (Freeman et al., 2014). In contrast, ALE have the benefit of increasing exam performance (Freeman et al., 2014), and information literacy (Detlor, Booker, Serenko and Julien, 2012). Even though the most frequently observed instructor behaviour in this chemistry course was presenting content or solutions, this only made up 35% of class time. Students collectively spend more time responding to instructor questions, working in groups and working on individual tasks than they did listening, especially in the workshop and laboratory class components. Classroom environments involving group work and case-based learning have been shown to benefit student motivation and self-regulation and performance, as well as improving the quality of educator-student and student-student interactions (Atwa, Gauci-Mansour, Thomson, and Hegazi, 2019). These instructor and student behaviours in this chemistry course are consistent with those in ALE. These observations made in this classroom, in light of students CT performance support the idea that engagement with scientific methods in ALE leads to CT development.

Aside from the teaching techniques, another reason this semester-long course was able to significantly increase CT performance is likely because there were multiple opportunities for students to practice the skills they were being taught. This was planned - as indicated in the alignment of learning outcomes and course components, as well as being evident in the classroom observations (COPUS data) and survey data. The benefit of ongoing practice is supported by evidence in the literature (Bransford and Stein, 1993; Willingham, 2002), with techniques such as problem-based learning being particularly effective (Leming, 2016). This course provided numerous opportunities for students to engage with scientific data and concepts and understand how an evidence-based conclusion is drawn. This likely explains why the course was able to contribute to a significant shift in performance on two questions (Q5 and Q11) that tested students understanding of using scientific data supporting a statement. These findings highlight two key aspects about CT development – the need to be intentional with learning opportunities (i.e. plan them and align them with course components), and the need to provide multiple opportunities for practicing CT.

The last issue to consider is that not all the observed gains might be directly attributable to the course. Question 5 required students to draw a conclusion; however, it was a closed question format requiring a yes or no response. These kinds of questions require less CT than open-ended questions (Husain, Bais, Hussain, and Samad, 2012), or like MCQ are just viewed as less authentic measures (Benjamin, 2012). The change in performance on one of the questions (Q10) is likely a combination of this increased understanding of using data to draw scientific conclusions and also some recall. Q10 is meant to be completed prior to opening the additional packet of information. However, the pre-test and post-test were identical, so there is the potential for students to recall the contents of the articles in the post-test even though the CAT creators suggested there are no learning effects (Centre for Assessment and Improvement on Learning, 2013). That being said the increase in students CT performance was greatest for Question 11. As the only question of the three significant results that required a written answer, this was the best indicator of improvement in student skills relating to problem-solving, communication, evaluation and interpretation of information (the collective groups of skills the CAT measures). Given that Australian policymakers promote an infusion/embedded approach, this finding of increased CT development is promising because it demonstrates that engagement with general scientific approaches (where CT content is purely embedded) can have a positive influence on CT development. This finding also lends weight to the findings from the Australian educator perception survey, more specifically to the science educator's belief that engagement with the scientific method can be effective for increasing CT development (Chapter 3, Section 3.4.2: Trends for question 4). However, it also begs the question of what more could be achieved if CT was incorporated in a slightly more explicit way?

The student survey results indicated that students were more likely to recognise being given the opportunity to practice general science skills than CT specific ones. For example, student in this course related/interpreted the experience of mathematically testing a claim using their chemistry skills as equivalent to determining whether an inference was supported by an advertisement even though this skill was not technically practiced in this course (in the sense of the intended meaning from the CAT instrument). The language of CT is often excluded from dialogue in science classrooms, leaving students to interpret classroom experiences in their own way. The lack of recognition of practicing skills could be because of the use of embedded CT development rather than explicit or mixed teaching methods. Being explicit has been previously found to have a positive effect on chemistry students CT development (Gupta et al., 2015). For example, in this study there was an emphasis on scientific writing and reflection (i.e. being more explicit about CT abilities), and there were significant gains in students CT. This finding is in line with other evidence in the literature which support the general benefits of explicit and mixed teaching methods for CT development (see Abrami et al., 2015). But perhaps the most important reason for Australian chemistry educators to be more explicit about learning outcomes including CT is to help educators and students gain a shared perception about CT. In an Australian study by Danczak et al. (2017) chemistry student and educator understanding and perceptions about CT in chemistry were found to be different. This reinforces Jacobs call for being more explicit about the CT to chemistry students (Jacobs, 2004), because non-content goals are less evident in embedded environments. In addition, Erikson and Erikson (2018) have suggested that there are some "educational goals that cannot expressed through learning outcomes" and that CT is one such example (2018, p.1). This means it would to be sufficient to claim CT include the learning outcomes if CT is not going to be explicitly mention and scaffolded into the course. The ACT Framework has the potential to help educators incorporate CT in their course. It provides a mechanistic way to explain CT the components and process of CT generally, as well as the links between these general skills and the expectations of what it means to be a chemist who can think critically.

There are other important benefits to being explicit. Being explicit about the relevance and context of principles in chemistry has been found to help with student success and motivation. For example, Ramsden (1997) explored the difference between 4 traditional science courses and 4 Salter's (context-based) science courses and noted little difference existed between the groups in terms of the level of understanding, but context-based approach (Salter's) stimulated students' interest in science more than the traditional science course approach. Similarly, other studies including a meta-analysis (Bennett, Hogarth and Lubben, 2003) found that context-based learning positively effects students interests, "attitudes … motivation and success" in science (Ulusoy and Onen, 2014, p.817). In 2009, the National Academy of Science also expressed that student context-based approaches which connect the subject to everyday experiences could possibly change student perceptions and chemistry achievement (National Academy of Science, 2009). Thus, to get chemistry students to experience more changes relating to the affective domain of CT then educators should consider being more explicit about CT and its relevance to chemistry/science.

4.7.3 Future learning opportunities in this course

Aside from generating broader understanding about CT development in chemistry, this case study process has also revealed some specific things about this course which could be targeted for improved CT development in future cohorts. It has also revealed opportunities to further develop some general science skills which also relate to CT.

In terms of general science skills - the majority of students did not think they got to practice the skill designing and planning an investigation, nor exploring the role between science and society. There was only one opportunity for students to design their own investigation, so this perspective was understandable. While one of the learning outcomes related to connecting across boundaries, the focus of the course was foundational chemistry, not chemistry and society. This meant connections were made between chemistry, physics and biology, rather than between chemistry and society. Examples did draw on student's everyday life; however, because they were linked to explaining a principle, students were likely focussed on the content rather than the application. Irrespective of the cause for student's belief about the lack of practice for certain skills, these represent opportunities to increase student engagement in divergent thinking, reflective thinking and problem solving, which in turn will benefit their capacity to do CT. The benefit of incorporating activities around designing and planning an investigation, and the role between science and society would be to prepare students to be better scientists. This is particularly important during a student's first year of study, as these experiences can significantly influence their study paths and therefore career choices (Baik, Naylor, Arkoudis and Dabrowski, 2019).

This case study process has also revealed a number of opportunities for improving CT development. Firstly, there were additional opportunities for CT development, as the majority of improvement was limited to one particular CT skill. With students achieving scores of less than

50% on Q3, Q4, Q6, Q7, Q9, Q14, and Q15 these represent good opportunities to start to increase CT development. All but one of these questions (Q14) also required some divergent thinking to generate alternative possibilities in their response. This was a skill which was practiced verbally in class, and was modelled by the instructors; however, these approaches did not translate into student responses on the CAT.

One of these questions required evaluation and interpretation (Q14), and four of these questions (Q4, Q7, Q14, Q15) required some element of problem solving. These are skill students practiced often in class; therefore, it is likely that the failure to perform better on these questions is related to the explanation and communication aspects of the mark scheme (see Chapter 6, Section 6.4.2 for criticism of methods). However, if the instructors were concerned about student performance of these skills then they could incorporate more problem solving practice in class, and/or seek verification of this in future cohorts using a CT test with subscales (such as the Watson-Glaser Critical thinking test, or the California Critical Thinking skills test) that can assess this skill in isolation.

All of these questions tested student communication skills, requiring explanations to support the answers whereas only one of the three questions where a significant increase was observed included this broad skill. This was a skill that students practiced verbally in class, but not in their written work. Therefore, this could easily be remedied by including at least one open-ended question requiring student to explain and support their thinking in the fortnightly quiz. The themes and scenarios for these open-ended questions could be made specific to demonstrate CT in chemistry; which could help to remedy both CT skills and increase student exposure to what CT looks like in chemistry in order to help overcome what Danczak et al. (2017) describe as a lack of shared understanding about CT in chemistry. The format of this open-ended question could be modelled on the CAT (See Appendix F) or the CLA (Klein et al., 2007); as it has been suggested that teaching to the test in these instances also results in development of the desired CT skills and competencies (Centre for Assessment and Improvement on Learning, 2013; Benjamin et al., 2014). These questions could help to make CT more explicit in the course. But given the otherwise embedded techniques, it would also be worthwhile introducing some explanations about CT using general language, with some scaffolding to link it to the relevant embedded activities. This could be achieved by introducing student to the ACT Framework, or another definition of CT as deemed relevant to CT in chemistry by the instructor.

Another significant lesson is the importance of including a CT assessment even in embedded environments. Had it not been for the CAT, there would have been no assessment of CT skills in this course. Findings by Leming (2016) emphasise the importance of aligning course assessments with targeted CT skills to improve student gains in CT. One solution for this, if it is not practical to continue to employ a CT assessment, is that the ACT Framework can be used to judge student work samples to start to determine where there might be flaws in their CT process. This would involve looking for evidence of CT process stages in student learning and then inferring where their critical thinking processes have not been adequately deployed (process stages based on the ACT Framework presented in Chapter 2).

Lastly, the fact this course capable of significantly improving CT when the evidence suggests embedded is the weakest teaching approach, is indicative that even greater gains could be achieved if CT was included in a more explicit way. Course redevelopment can be a time-consuming exercise, but because CT has also been found to be a predictor of success in chemistry (Fredette, 2018) the payoff for investing more time in CT in chemistry courses could be higher quality chemistry graduates. However, given that the findings from this study did not verify this; further research needs to be done.

4.8 Conclusions

Through reflecting on class learning opportunities and on CT performance and perception changes this first case study has contributed to understanding about the types of teaching approaches used in first year chemistry and their effect on CT development. In doing so it has started to address the second broad research question for this dissertation - *What teaching approaches are currently employed by science educators in order to develop critical thinking abilities, and which are most effective?* This chapter is the first of three which contribute understanding of CT development in science. It has provided a method to explore CT development using a trio of measures that provide understanding into experiences and opportunities that make up the classroom ecosystem and the learning that results from them. These measures have helped reveal the actions and outcomes from the existing chemistry course, as well as highlighting opportunities where CT development could be improved. It sets the scene for Chapter 6, which further demonstrates the capacity for this method, in addition to the ACT Framework, to inform instructional change.

This chapter has contributed understanding to objective 2, revealing that ALE which explore and apply scientific principles can have a positive effect on students CT skills. Ultimately this showed that it is possible for a discipline-specific science course with embedded CT learning opportunities to produce a significant improvement in CT skill performance. But despite the fact that the course was effective at developing CT, findings from the literature indicate that it could have been more so if the course incorporated a mixed approach to CT development (general plus infusion or immersion). Overall, the magnitude of the performance change in line with the use of approaches which have previously been associated with CT development support the belief that this chemistry course did have a positive effect on CT development. These findings increase and enhance understanding about Australian science educators' perceptions about CT, and contribute evidence about the effectiveness of certain teaching methods used in science.

Chapter 5

Exploring Critical Thinking Development in Biology and Society

The saddest aspect of life right now is that science gathers knowledge faster than society gathers wisdom.

Asimov and Shuman, 1988, p.281

5.1 Introduction and chapter outline

There is a growing need to train undergraduate scientists to consider the multi-faceted social and ethical perspectives that emerge through scientific advancement (Lautensach and Lautensach, 2010; Almeida and Quintanilha, 2017). The nexus between science, society and education is not new, with calls for knowledge sharing among these spaces being promulgated at various points in the last 40 years (see Ziman 1980; Cetto, Schneegans and Moore, 2000; Lakomý, Hlavová and Machackova, 2019). However, the rapid pace of data generation in modern society (Helbing et al., 2019) amid student's tendency to rely on Google and Wikipedia for information (Head, 2013) adds further pressure to the need to equip students with scientific literacy skills. Science and society courses are one solution to address this issue, as these courses are specifically designed to emphasise the importance of thinking critically about the different ways scientific messages and opinions are conveyed to society. These courses are particularly beneficial because communication, argument analysis and reflective thinking remain central priorities to the courses, without these scientific literacy skills being lost among other more content-based learning outcomes (Rutledge, Bonner and Lampley, 2015). Science and society courses are especially vital amid the growing calls to contextualise science for students (Llopart, and Esteban-Guitart, 2017; Mouton and Archer, 2018). Therefore, it is timely to determine how effective science and society courses are at developing student's critical thinking (CT) skills.

The effects of science and society courses has been previously explored in the US (Gottesman and Hoskins, 2013; Rowe et al., 2015). In both instances, a significant increase in CT performance was measured using the CAT instrument, with between 23-32% overall score improvement achieved in a single semester. However, completion of a general education science course (such as science and society) is a requirement for many US degrees, whereas in Australia it is generally offered as an elective course. As a result, the function of the science and society courses is likely to be different in Australia, as science students should be trained in these communication and argument analysis skills within their core courses. However, there is increasing evidence that the embedded-style teaching is not equipping students with the scientific literacy and communication skills needed (Stevens, Mills and Kuchel, 2019). Therefore, gaining insights into the broader value of a generic communication course for science students, including their effect on student CT development, will lend more support to arguments for these types of courses being included as a core component of bachelor degrees.

This chapter forms the second of a three-part story on the development and assessment of CT in tertiary science classrooms. The specific focus is student CT development in an Australian science and society course that is taught at Flinders University. The course, biology and society, explores current biological issues, exposing students to perspectives that come from both experts and society. Given previous findings about science and society courses in the USA, as well as the structure of this course which includes interactive discussions and assessments requiring the use of logic and argument, it was anticipated that this course would improve students CT skills. However, findings from the first data set revealed that the effects of the course were negligible. As a consequence, some changes and interventions were incorporated into the course. A second semester of data was undertaken to explore the effects of these changes and a positive shift in student CT abilities was observed.

This chapter addresses research aims 2 and 3 (see page 9) and continues to contribute to understanding about CT development in tertiary science in Australia. It builds on the chemistry case-study chapter findings by further demonstrating the value of the multi-tool evaluation approach (in combination with my ACT Framework) to inform meaningful pedagogical change. An important difference about this chapter compared to the previous chapter is that this component of research models the diagnostic and improvement process that was foreshadowed in Chapter 4. Specifically, the classroom observation data highlighted that the biology and society learning environment was more passive than active, and there were limited opportunities to individually practice CT even though the course mentioned CT often. The CT performance results also revealed specific weaknesses in student performance. This information allowed for some evidence-informed changes to be made to the next version of the course. Thus, this chapter makes both methodological contributions, as well as knowledge advancements, by exploring CT development in an Australian science and society course.

5.2 Biology and society case study method

Research tools outlined in Chapter 4 (Exploring CT Development in Chemistry Section 4.4) were also applied to the biology and society case study, with some modifications outlined below. Again, the course evaluation process consisted of a pre-test post-test method in a quasi-experimental design; however, for this case study two semesters of data were collected. In both semesters, observations of the biology and society (B&S) course experience commenced after the students had completed the first CAT test in the second lecture of week 1 and ended upon completion of the CAT in the final session of the course during week 12. There were a few differences which arose from logistical considerations. For example, the CT test and student perception survey were both completed on paper because of the type of classroom data was collected in (a lecture theatre). There were also some differences given the fact that version 1 of the biology and society course (B&S-V1) was the first course data was collected in. Essentially, this first course served as a pilot for the case-study method, in particular the student perception survey, with subsequent deployments of the case study approach incorporating changes to the student survey (as noted in Chapter 4 Section 4.4 and Appendix D).

Another important difference for this course was that data was collected over two consecutive deliveries of the course, with the second version including four intervention tasks. Each of these modifications are outlined below.

5.2.1 CAT protocol modifications

Due to logistical considerations around ethics approval, permission to use the pre-test Critical-thinking Assessment Test (CAT) data from the B&S-V1 cohort was sought with a permission slip distributed with the post-test. Scoring of the CT performance measure was undertaken myself and the B&S course coordinator (in other words, different scorers from the chemistry case study), with a third perspective sought from a postdoc education researcher in the few instances when the assigned scores did not agree.

5.2.2 Perception survey modifications

The student perception survey was always designed as a pre-test post-test capture system. However, delays in ethics approval meant the perception survey used in B&S-V1 could only be distributed as a post-test. To compensate for this, adjustments to the B&S-V1 post-test wording were designed to capture the same understanding (of student perception of their improvements on each skill). However, upon seeing the initial survey results it became evident that using a pre-test post-test survey was definitely needed to capture changes in student perceptions of their skills. This is why chemistry and B&S-V2 had a pre-test and post-test perception survey, and B&S-V1 did not (Figure 5.1). An additional change that resulted from the B&S-V1 and chemistry results was the addition of more disposition items (Figure 5.1; see also Appendix D: Table D.1). Literature suggests that dispositions and the willingness to think critically can change before cognitive CT skills (Bowman, 2010; Culver, Braxton and Pascarella, 2019). However, better outcomes can be achieved when the intellectual characteristics of being a critical thinker are incorporated in a course alongside skill practice (Facione, 2000; Kuhn, 2019). The addition of more disposition items in B&S-V2 was intended to help clarify the contrast between B&S students' perceptions of the skills they thought the course had helped them improve and their CT performance results. Was it a failure to deploy skills, or was it a lack of CT abilities? These item variations are noted in Appendix D. Content was kept consistent between the pre-test and post-test used within each course.

5.2.3 Classroom observation modifications

There was a small change to the aspects of EQUIP instrument used in this case study compared to chemistry (Section 4.4.1). Specifically, two of the instruction factors (teacher role, and the student role) were also incorporated into judgements of the inquiry level to help assess the types of student learning opportunities offered in the course (See Appendix E: Table E.4). These aspects were meant to be captured through the other observation instruments, however, the variability in instructors in the B&S course (>20 different presenters compared to three in chemistry) meant that the summaries generated by the COPUS instrument and QSRLS instrument did not capture the extent to which the different presenters and teaching sessions enabled student to be the masters of their own learning.

5.2.4 Changes to B&S-V2 course

Large improvements were expected in B&S-V1 because of the results revealed in previous studies of science and society courses in the US (see Gottesman and Hoskins, 2013; Rowe et al., 2015), and also because of the use of explicit discussion about CT in class (Abrami et al., 2008 2015). However, there was limited evidence of improvement to CT development in the B&S-V1 course (see Section 5.4). As a result, pedagogical changes and three intervention tasks were introduced in the B&S course to specifically advance CT skill development. These pedagogical changes included reducing the number of documentaries watched in class (passive learning task) to increasing the amount of in-class practice of CT-related abilities (active learning tasks) (see Section 5.3 for additional information). Importantly, these changes were informed by the evidence gathered using the multi-tool assessment process. For example, the pedagogical changes were informed by the observation data summaries, which revealed a mostly passive learning environment, so while CT principles were discussed there were not many opportunities to practice these skills. The CAT results also indicated that more time needed to be spent considering appropriate forms of evidence and alternative explanation to claims, as students performed poorly on these questions despite these ideas underpinning expectations in students' written work. The

CAT results also helped identify changes in student CT performance. This enabled improvement efforts to be focussed on areas of weakness in the context of CT skills that were important to the overall course learning outcomes (see Section 5.4.3). More information about the intervention development is provided next.

CT skill intervention task development

As mentioned in Chapter 4, part of the training for the CAT involved learning how to develop CAT-like questions to help develop student's CT skills. This training was implemented in the B&S-V2 course because students CT performance in the B&S-V1 course was measured to decline rather than increase. The development of interventions was a collaborative effort between me and the course coordinator (CC). Prior to intervention development, a review of the course was undertaken. The first step in this process was identifying the CT skills to target for improvement - finding the balance between course goals, learning outcomes and CAT measurable skills; and also figuring out a suitable theme for scenario development. CAT results were examined on a question by question basis, with performance compared against the intended course learning outcomes. Next consideration was given to the learning opportunities (location and experiences) associated with the CT skills that the course was expected to develop. An early version of the Adaptive Critical Thinking Framework was also used to help explore the aspects of CT present in the course and its assessments. Given the course's learning outcome "appreciate how evidence is used in a scientific argument" (Hunter, 2015, p.2) particular focus was given to the framing of evidence and argument in the course. Areas of weakness were identified, and course themes were explored to identify potential opportunities for CAT-like question scenario development. Decision-making about the CT skills to target and lecture content related to the interventions was shared, scenario development was the researcher's responsibility and scenario delivery was conducted by the CC.

The specific skills identified for intervention related to 'recognising the strength and flaws in an argument' and 'identifying and creating solutions to real-world problems;' with the subject matter of week 5 (genetically modified organisms), week 8 (viruses and bacteria) and week 11 (pregnancy) providing the best opportunities for scenario development. This spread of CT skill practice across the semester is also supported by recommendations in the literature about providing multiple opportunities to practice skills to increase learning (Leming, 2016; Svihla, Wester and Linn, 2018).

Scenarios were developed in line with CAT questions 5, 10, 11, 13, 15 (see the list of skills associated with these in Appendix J.1). Skills from other questions (2, 3, 4, 6, 7, and 9) were also incorporated into class discussion and group activities, as they were relevant to course learning outcomes. Whilst providing opportunities to practice CAT-like questions could be seen as training to the test, the type of training is not about gaming the system to produce a stronger performance for metrics related to national or global standings. These opportunities are designed to train students to be better critical thinkers, so some authors believe the ends justify the means if "we would be teaching the competencies we want to develop in students" (Benjamin, 2014 p31). In the case of this course, scientific literacy and argumentation form a major part of the course assessment, and as such, CT skills are needed to be able to gather relevant information, analyse and synthesise ideas and make evidence-based judgements. Therefore, training students to be better critical thinkers using CAT-like questions not only helps on the CAT test (which was not associated with their grades) but it would help them develop their science literacy and argumentation skills needed to be successful in their studies and lives. However, as a precaution to avoid the association between the intervention tasks and the CAT test, the interventions were decontextualised. Students were provided with a mix of written and verbal response questions relevant to the theme of course content, incorporated through active learning strategies such as individual tasks and group activities. This combination ensured that the interventions fitted in with the typical activity style of the B&S course and did not stand out as training students to complete the written CAT assessment. Written interventions requiring individual participation were used in week 5 and 8 and 11. The week 5 intervention also included a scaffolded discussion with a follow up individual question, and the week 8 intervention also incorporated a verbal group response. Specific information about these tasks can be found in Appendix J.

5.2.5 Data analysis

Statistical analysis

The initial analysis process was the same as the methods in the chemistry case study (Section 4.4.2), where the outcomes from B&S-V1 and B&S-V2 were analysed separately to identify key trends within each course version. However, additional analyses were undertaken because this study also compared the difference in outcomes from two versions to reveal the effect of the pedagogical changes and interventions. This included some quantitative comparisons of the two versions of the magnitude of change in student CT abilities between course, as well as the inclusion of PCoA ordinations of COPUS data to compare the difference in class delivery as a means to help explain variation in student CT development.

To compare between the classroom environments using multivariate analysis methods, additional data processing of the COPUS data was required. First, to ensure equal weighting between COPUS counts per teaching session, the data was transformed into proportions of each code per session. This meant a longer class session did not carry more influence on the distribution than a shorter one. Next sample-similarity matrices using the Bray-Curtis similarity algorithm (Bray and Curtis, 1957) were generated for each course, as well as for the B&S-V1 and B&S-V2 comparison (Figure 5.1). For each of these data matrices, PCoA ordinations were generated. Further explorations of the course characteristics were undertaken through the unconstrained Pearson vector overlays, to see which of the COPUS codes moderately to highly correlated to the data cloud. In addition, to further elicit trends from interventions, comparisons of the change in CT performance each semester were undertaken using a Wilcoxon signed-rank test to gain deeper insight into the patterns across the two versions of the courses. A one-sample means test was also used to compare the overall CAT scores from B&S-V1 and B&S-V2 to the American freshman norm data. These multivariate analyses were performed using GraphPad Prism (v.7), and PRIMER (v.6) with the PERMANOVA add-on PRIMER-E, Plymouth Marine Laboratory, UK, (Clarke and Warwick, 2001). A summary of the statistical approach is shown in Figure 5.1.

Sample size

Cohort size varied from year 1 (n= 89) to year 2 (n=78), and data was collected from all students who attended the lecture in week 1 and week 12. However, after data processing the samples consisted of 30 participants from B&S-V1 and 37 participants from B&S-V2 (which equated to between 30 and 50% students enrolled in the course each year respectively). However, this participation rate was not considered a problem, as it was fairly reflective of the number of students who regularly engaged with the course (class attendance is taken by the CC including for students who only engage with the content online). It is also double the amount of paired data the CAT creators suggested was needed to be able to determine the course effect.

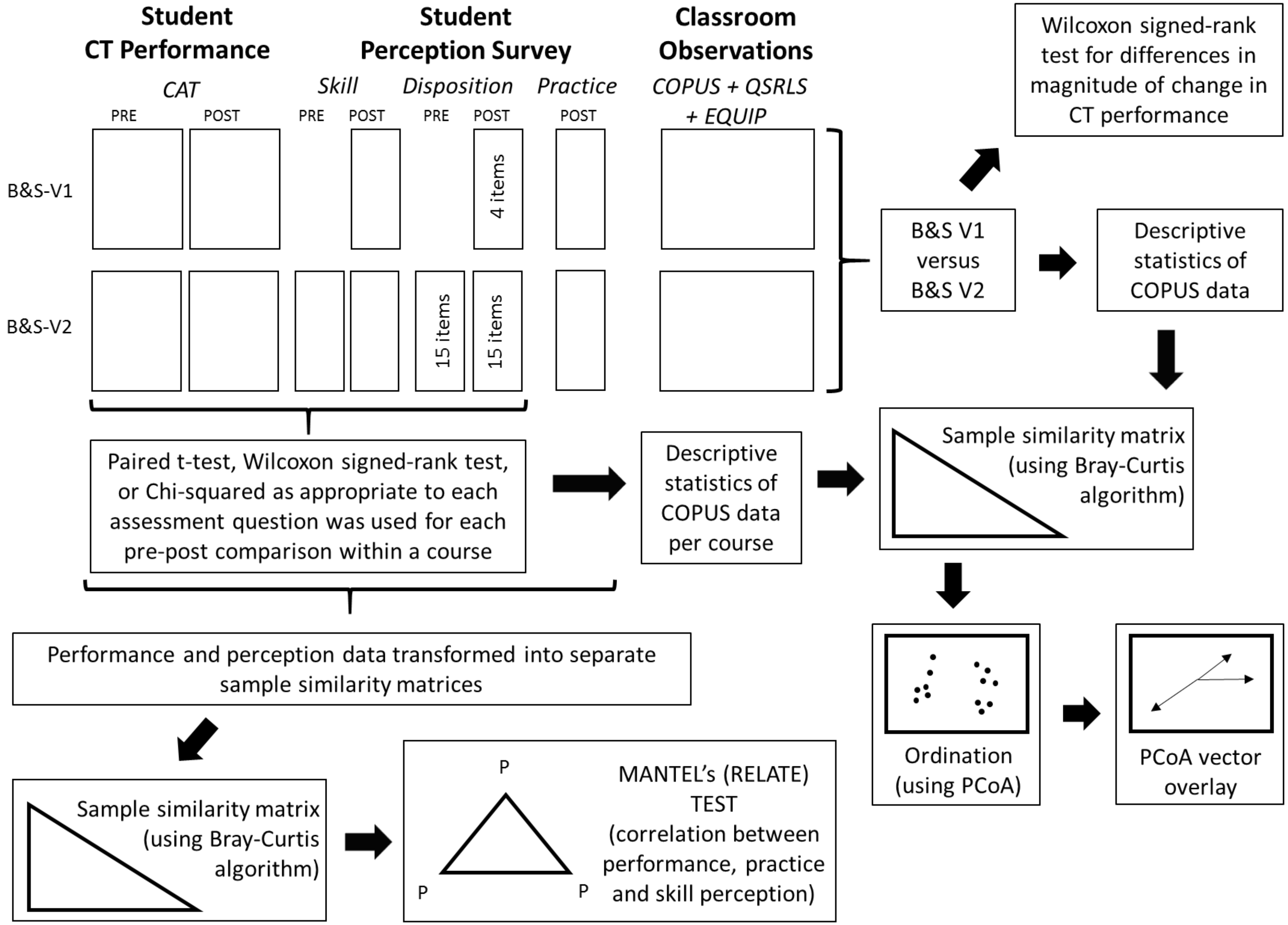


Figure 5.1 Summary of the statistical analyses undertaken within and across the two versions of the biology and society course.

5.3 First-year science and society course structure and background information

Biology and society is a semester-long elective course designed to expose and encourage students to explore and reflect on the evidence, ethics and issues that arise when science and society meet. The course is available to all undergraduate students at Flinders University but is typically taken by first-year students, many of whom study science. The aims of the course are to - "Learn how socially relevant science really is; how to critically analyse what you read, or see on T.V.; improve your writing skills"; and "how to have an educated opinion and influence people" (Hunter, 2014, p.2).

Class time consisted of lectures (65% of course time) and workshops (35% of course time). This breakdown was consistent between the B&S-V1 and B&S-V2 course versions (Figure 5.2). In both years the majority of lectures were delivered by guests who were experts in the theme for the week, with the workshop class being led by the course coordinator. Learning outcomes for the course were the same between the B&S-V1 and B&S-V2 course versions (Appendix I: Table I.1). The main change to B&S-V2 was the pedagogy associated with workshop delivery. For example, in B&S-V1 the workshop time consisted of watching documentaries with some discussion time; whereas for B&S-V2 workshop times incorporated far fewer documentary sessions with more time allocated to class discussions (Figure 5.2). In line with the increase in class discussions, there were also changes made to the delivery of the discussion sessions. In particular, there was a weekly emphasis on argument construction and how to assess and use evidence, coupled with both individual and group tasks that enable students to practice these skills. There was also greater emphasis placed on self-reflection, with the introduction of reflective thinking responses instead of a password to take class attendance. Three of these discussion sessions were also configured to include CAT-like questions to serve as CT development interventions. The process involved in the development of these interventions is discussed further in the next section.

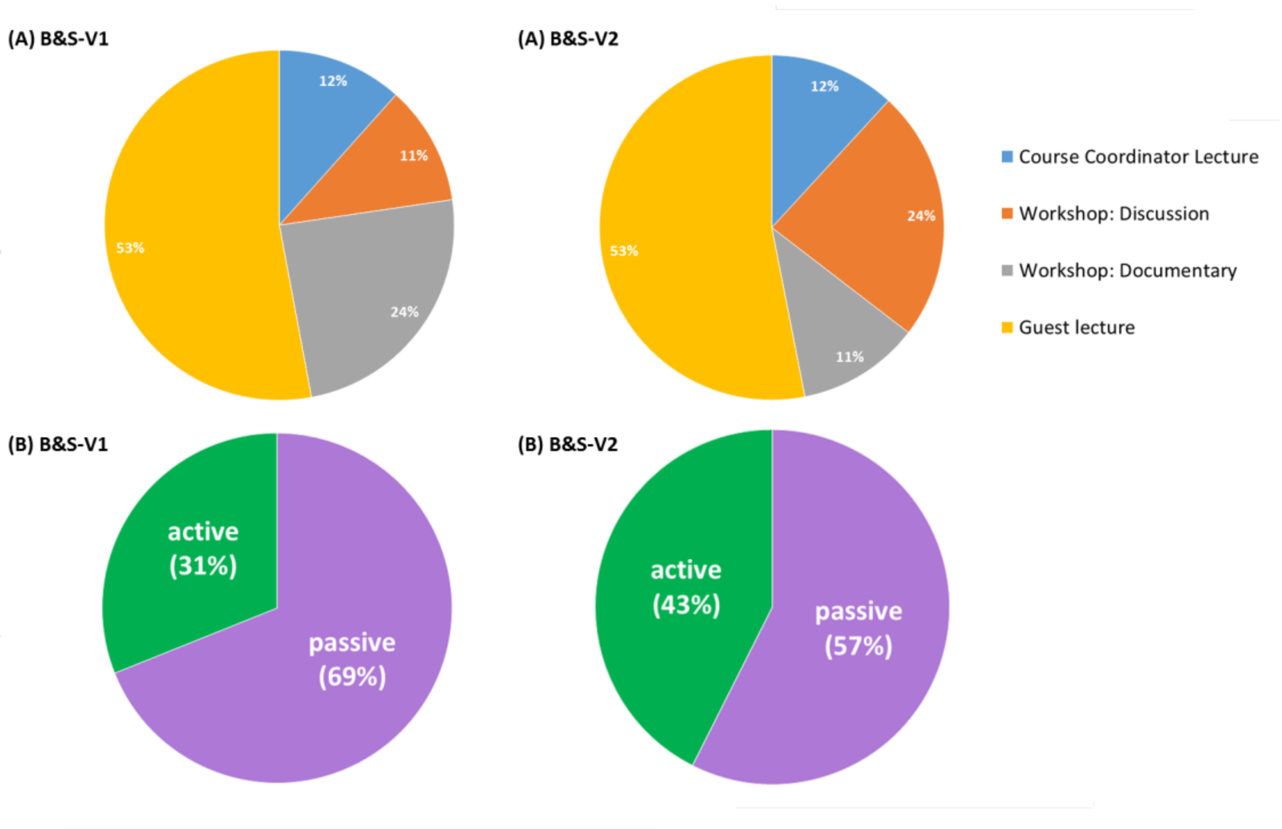


Figure 5.2 Summary of activities (A) and active learning time (B) for of the two versions of the biology and society course (B&S-V1 and B&S-V2).

5.4 The effects of a first-year science and society course on CT development

This case study facilitated exploration of the previous finding that critically analysing themes and arguments relating to science in society will produce a positive effect on CT development (Rowe et al., 2015; Gottesman and Hoskins, 2013). However, this study explores this course type in a new context - specifically, in an elective course in Australia where the majority of students taking the course happen to have a science background (even though the course is available to all students at Flinders University).

As demonstrated with the previous chapter, the results are divided into subsections which summarise the core findings pertaining to explorations of classroom observations; explorations of student perceptions about the course learning opportunities; explorations of student perceptions of their ability levels; and explorations of student CT performance. However, two semesters worth of data were collected on the B&S course. Therefore, one key difference between this chapter and the previous one is that the findings from the two semesters are presented together to facilitate easier comparison of the differences between learning opportunities and outcomes in the two versions of the course. In summary, overall student perceptions of their CT abilities significantly improved, but CT test results did not indicate a statistically significant increase in CT performance from pre-test to post-test in either version of the course.

5.4.1 Summary of classroom experiences

Observation tools findings: COPUS

Figure 5.3 and 5.4 are summaries of the student and instructor behaviours during the biology and society (B&S) course. In both course versions, instructor time was used effectively to engage students with course content, with the total sum of experiences comprising of >5% of 'other' non-content related activities. The most frequently observed behaviour in both iterations of the course was students listening (Figure 5.3A). However, with the changes made in B&S-V2 there was a 12% reduction in the amount of time student spent in this passive learning mode (refer back to Figure 5.2). This time was instead spent with students engaged in more active learning tasks, such as classroom discussions, answering instructor questions and working on individual problem-solving tasks. In line with this shift from a passive to a more active learning environment (ALE), there was a small change in instructor behaviours with 6% less time presenting content and 5% more questions (Figure 5.3B).

There were behavioural variations by instructor, and also by session type. For example, in the B&S-V1 course, the guest lecturers contributed the most to the amount of course time spent in passive modes, and discussion sessions the least (Figure 5.4). In both course versions the discussion sessions, run by the CC, included the broadest range of activities. The variation between the instructors and the session type was further evident when multivariate approaches were applied. The PCoA ordinations in Figure 5.5A and 5.6A demonstrate that there were differences between the way the CC and the guest lecturers presented, even when the CC was delivering a lecture rather than a discussion session. However, there were also differences between the way the CC presented depending on the type of session being taught. In both versions of the course, Pearson vector correlations indicate that the shape of the data cloud was most strongly influenced by class time being spent in more passive modes (such as presenting and listening format) versus more student-directed and active modes (such as students engaging in individual thinking or group work, and the degree of time spent in instructor-led discussion) (see Figure 5.5B and 5.6B). This trend is clearer in the B&S-V2 course, where there was more consistency in the CC's teaching style (Figure 5.6B). However, in both instances, guest lecturer presentations and documentary sessions cluster together (Figure 5.6A), and the CC presentations sit apart from this main cluster in two groups of their own (depending on whether the class was delivered as a lecture or a discussion session). Additionally, when viewing both courses together in multivariate space, there were also some trends evident in the sessions of the same type (Figure 5.7 A and B). For example, the lecture sessions in both versions of the course tended to cluster together (albeit there were clear differences between the delivery approaches of the guest lecturers and the CC). However, the shape of the left side of the data cloud indicated that the CC had changed their approach to discussion sessions between the B&S-V1 and B&S-V2 course versions.

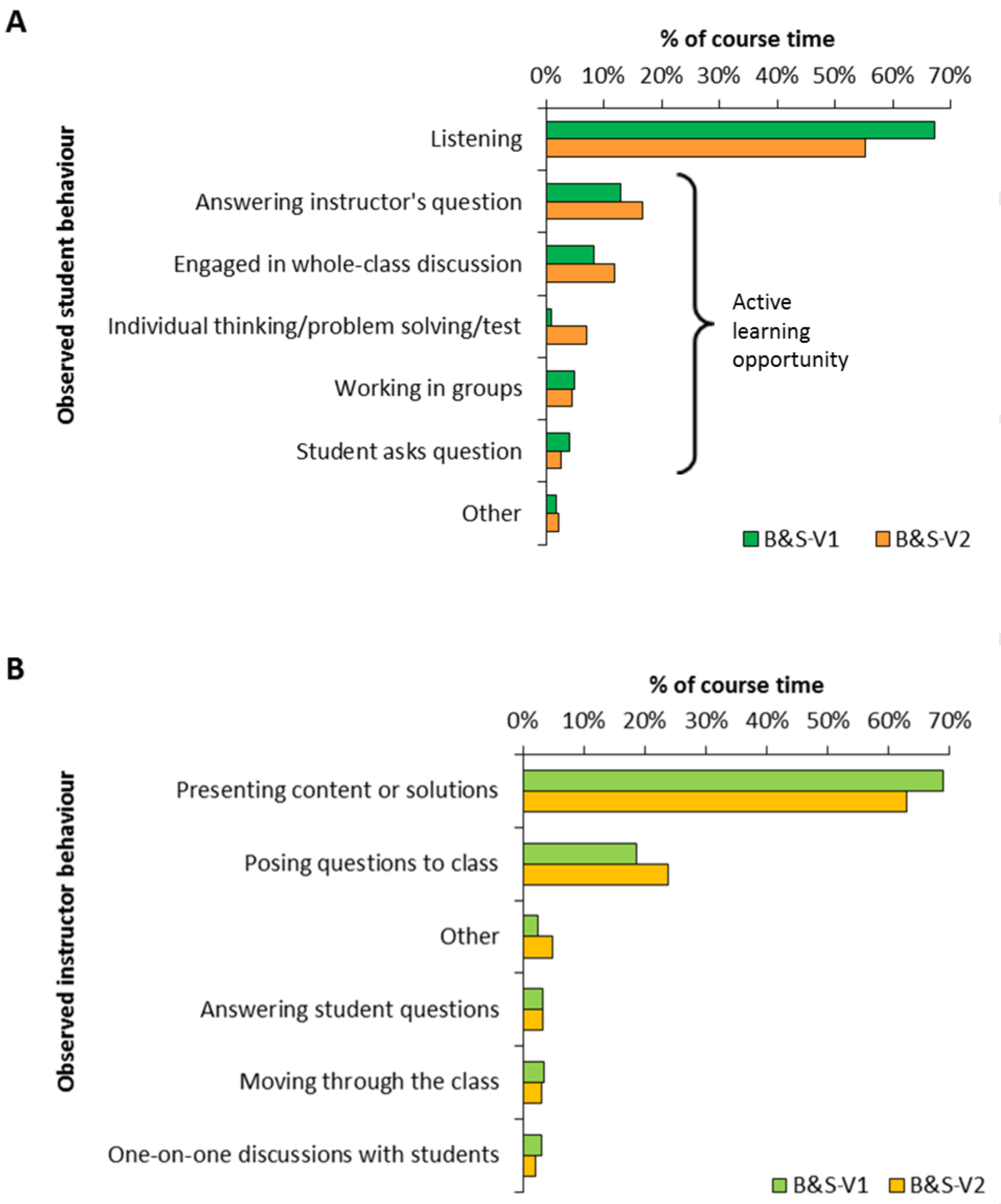


Figure 5.3 Summary of how class time was spent in the two versions of the biology and society course, where (A) compares how student time was spent between the two versions and (B) represents how instructor time varied between the two versions.

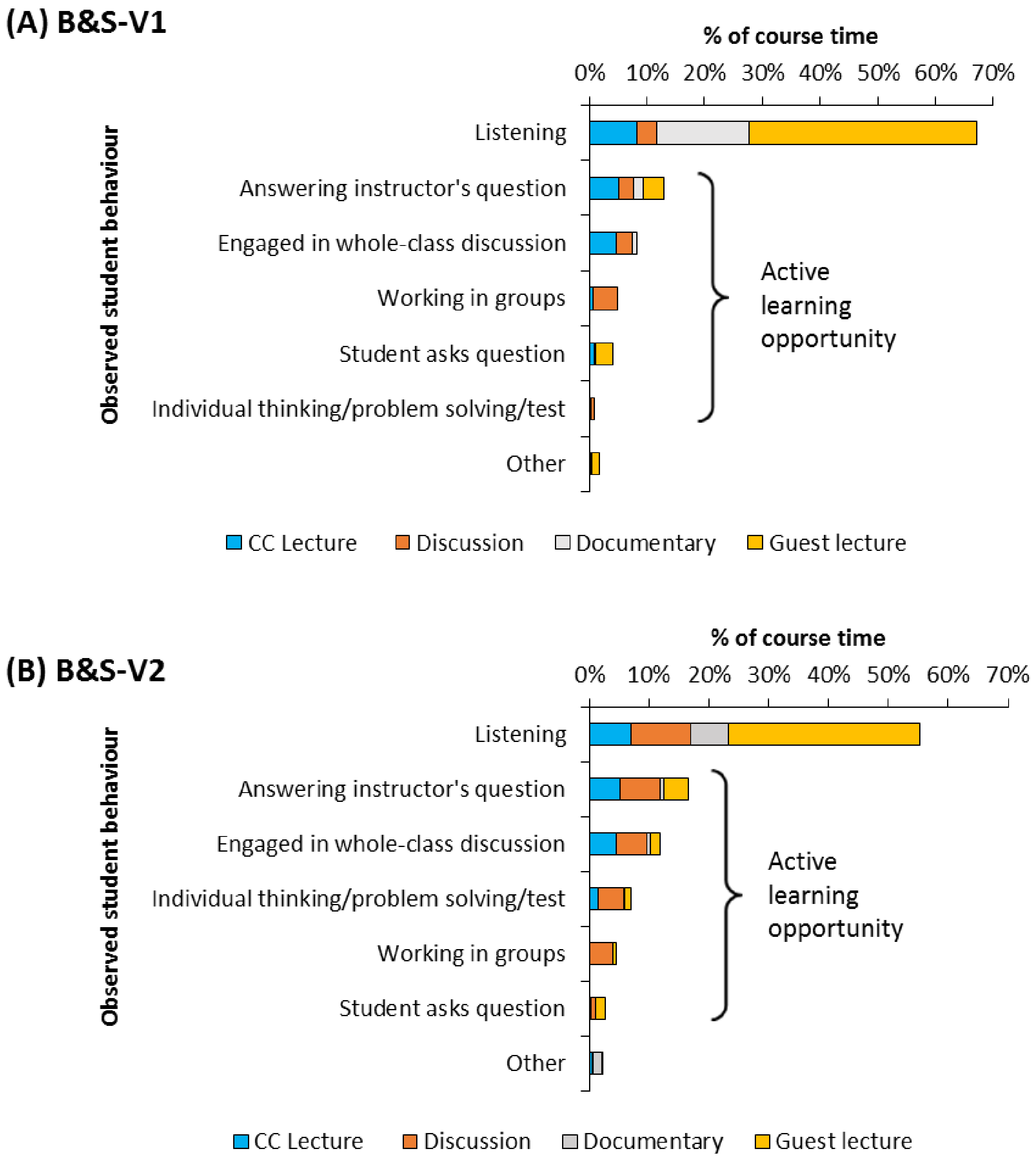


Figure 5.4 Summary of how student class time was spent in the two versions of the biology and society, where (A) shows B&S-V1 and (B) B&S-V2. Bar stacked to convey contributions by per session type. Summary of instructor time found in Appendix K: Figure K.1 and K.2

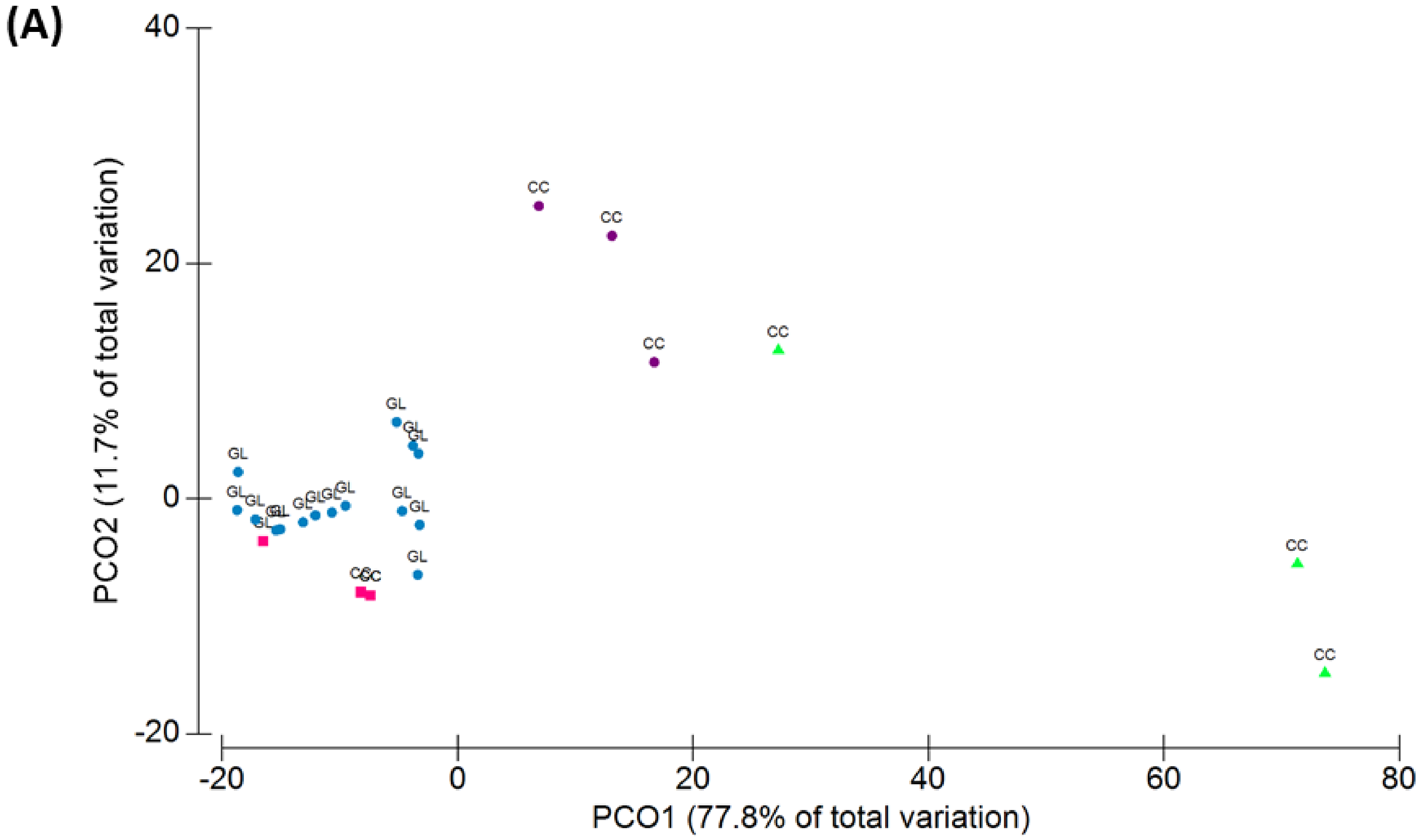


Figure 5.5 (A) Comparison of how each class time was spent in the B&S-V1 course using PCoA, which displays the variation in teaching within the course: pink squares = documentary; blue circles = guest lecture; green triangles = discussion; purple circles = course coordinator lecture.

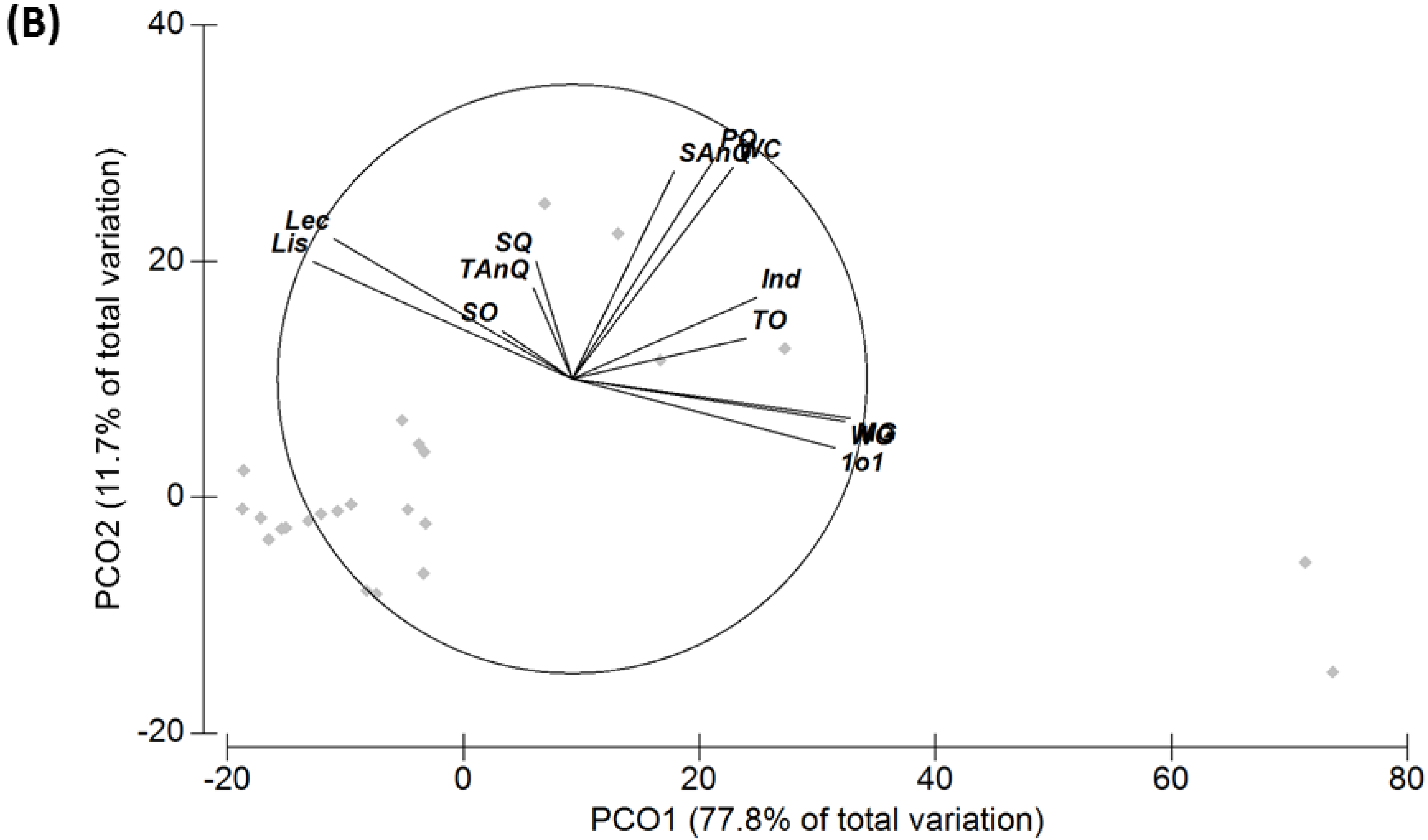


Figure 5.5 (B) Vector overlays on Figure 5.5 (A).Comparison of how each class time was spent in the B&S-V1 course using PCoA. The direction and length of the vector indicate the extent of the correlation between COPUS codes and the shape of the data cloud.

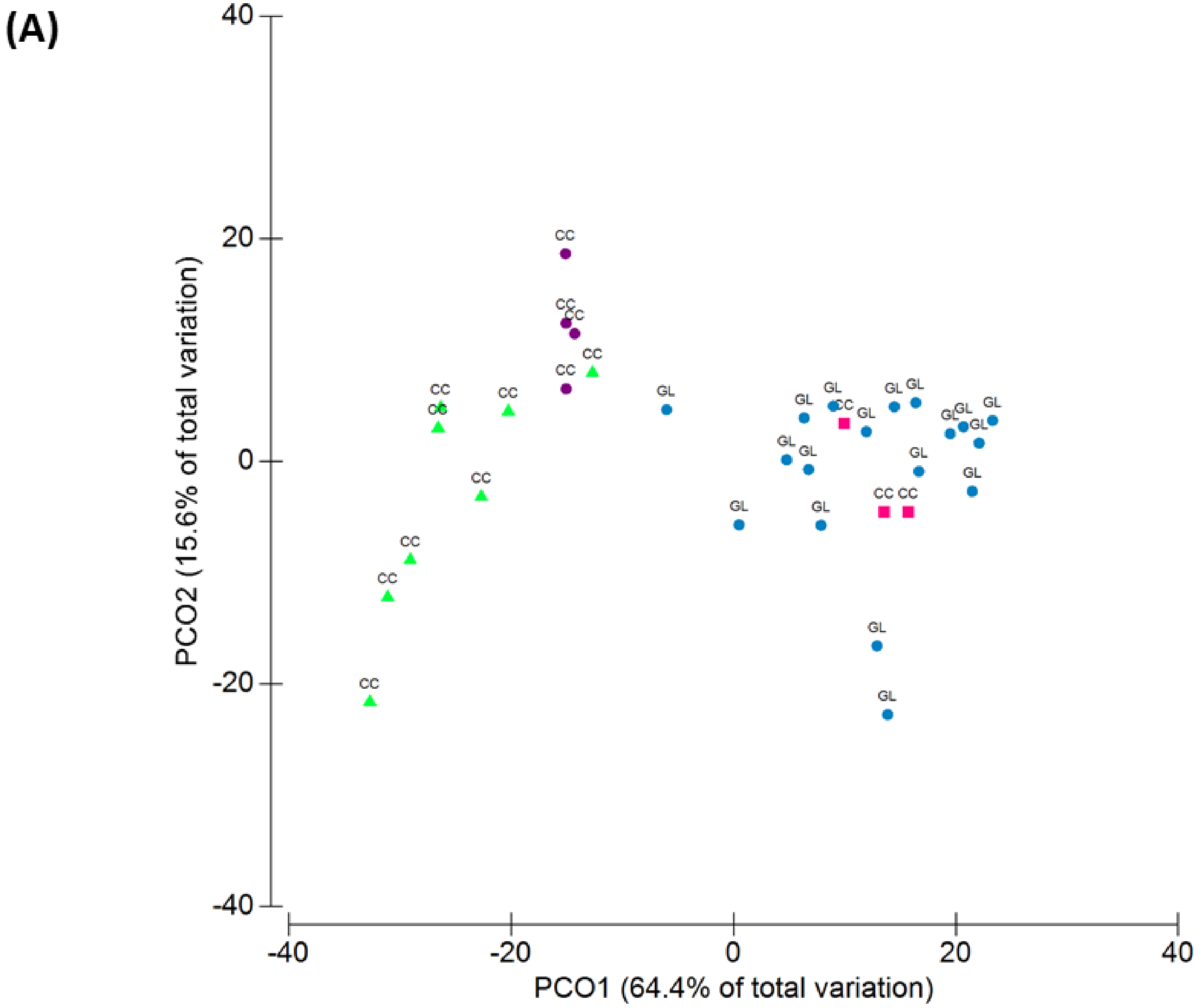


Figure 5.6 (A) Comparison of how each class time was spent in the B&S-V2 course using PCoA, displaying the variation of learning experiences within the course: pink squares = documentary; blue circles = guest lecture; green triangles = discussion; purple circles = course coordinator lecture.

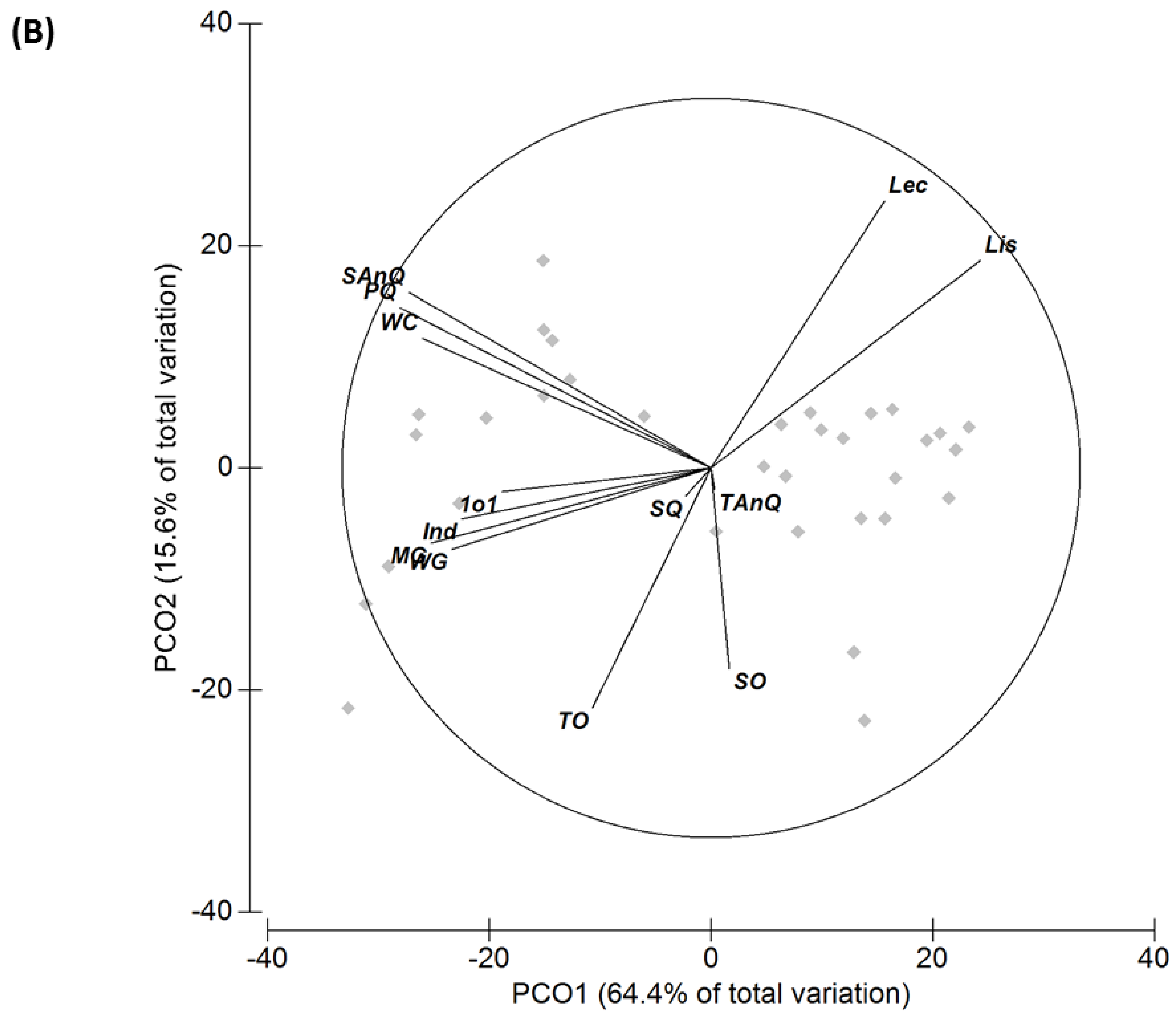


Figure 5.6 (B) Vector overlays on Figure 5.6 (A). Comparison of how each class time was spent in the B&S-V2 course using PCoA. The direction and length of the vector indicate the extent of the correlation between COPUS codes and the shape of the data cloud.

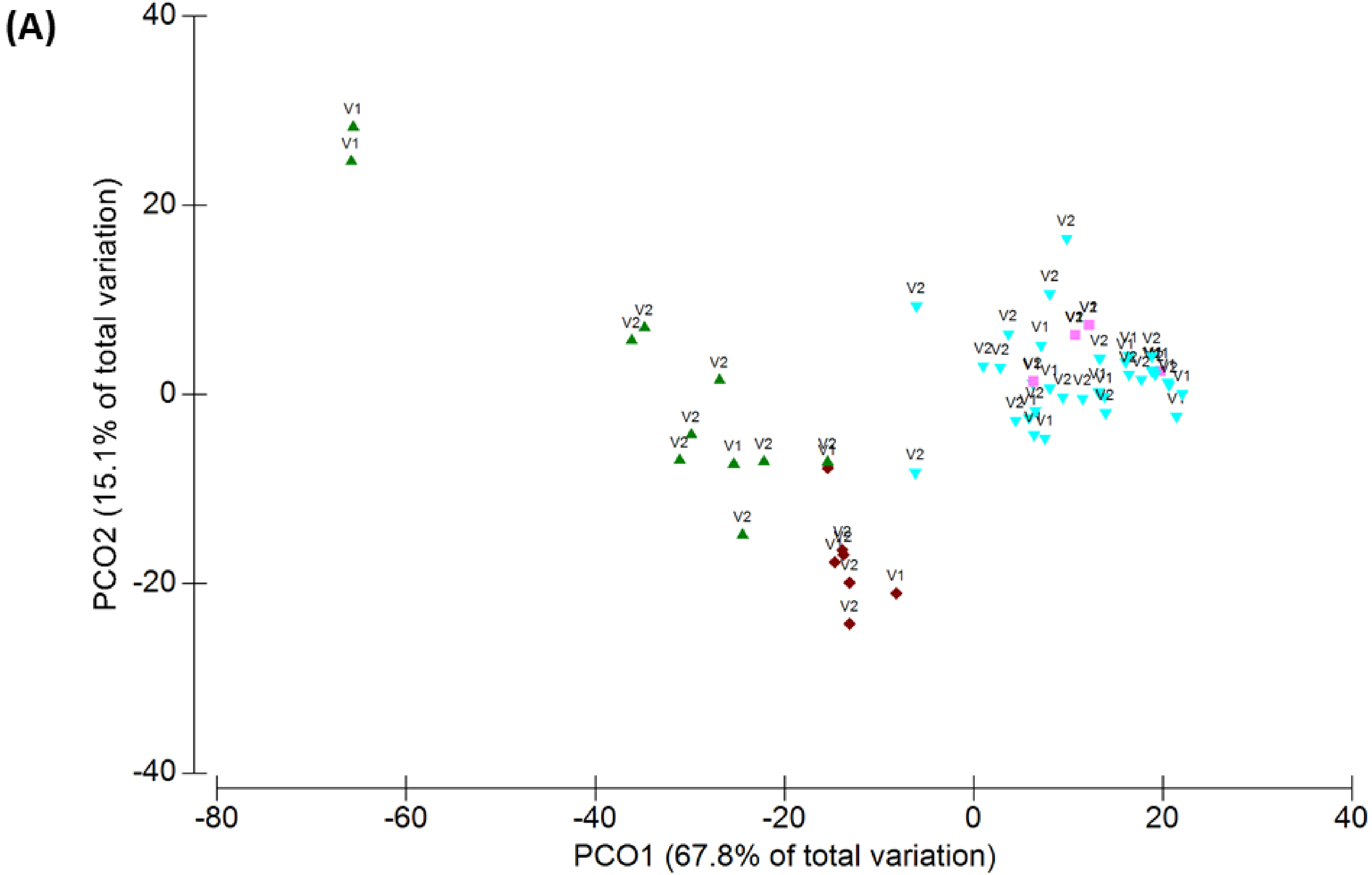


Figure 5.7 (A) Ordination of the various experiences students were exposed to in the two versions of the biology and society course, showing the spread of classroom experiences according to session type where pink squares = documentary; blue triangles = guest lecture; green triangles = discussion; purple diamonds = course coordinator lecture.

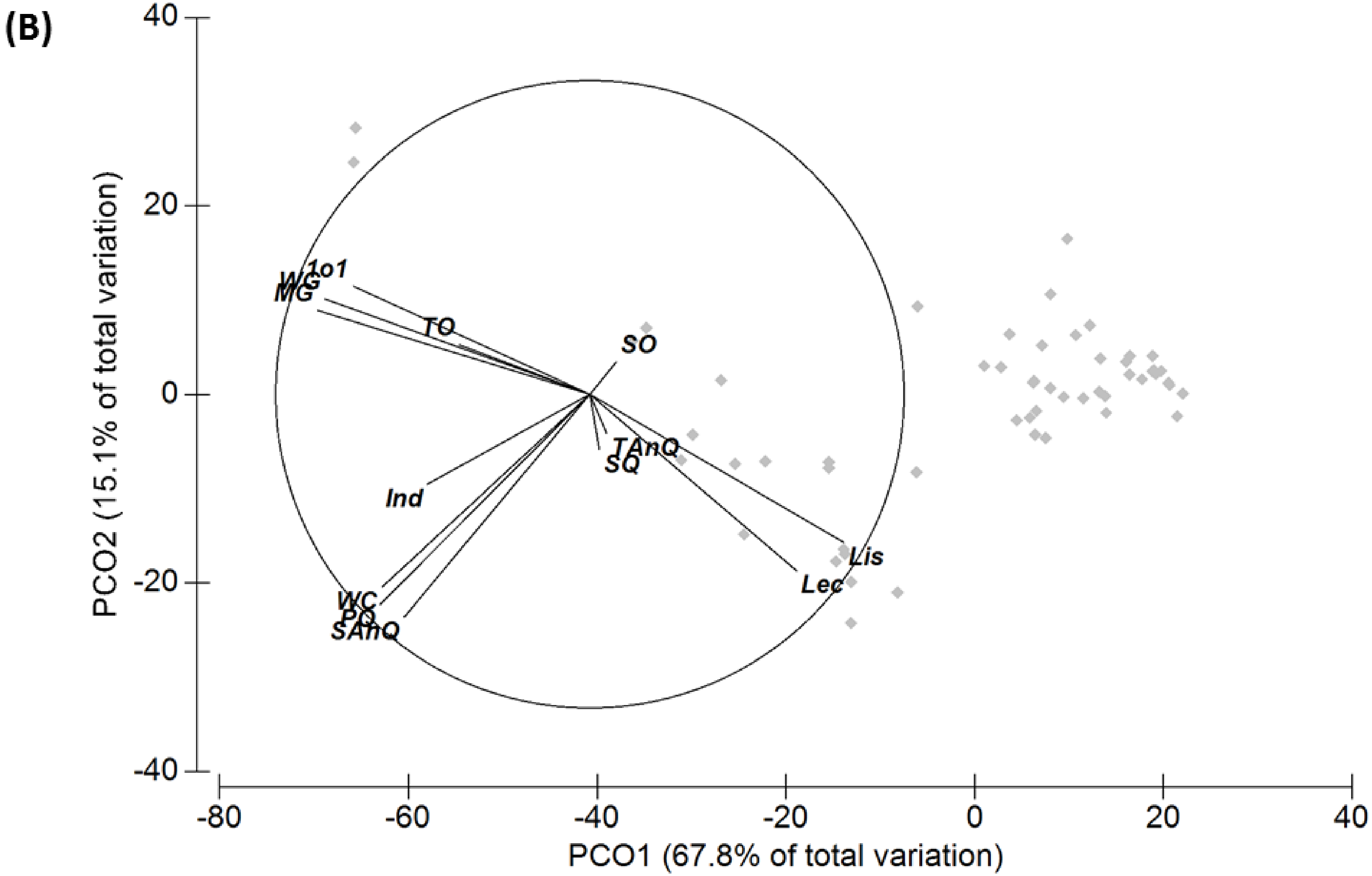


Figure 5.7 (B) Vector overlays on Figure 5.7 (A) the ordination of the various experiences students were exposed to in the two versions of the biology and society course. The length of the vector indicates extent of the correlation between COPUS codes and the classroom experience related to that portion of the data cloud.

Observation tools findings: QSRLS

Three aspects of the QSRLS protocol (knowledge integration, problematic knowledge, problem-based curriculum) were used to categorise how knowledge was treated in the classroom for both iterations of the B&S course. The scoring parameters for the QSRLS dimension 'knowledge integration' do not fit very neatly to this course because the purpose of the course (in conveying themes relating to science in society) was always going to produce a blend of science, social and cultural subject matter. After consideration of the fact that the thinking and communication skills the course focusses on are commonly discussed in terms of their broader applicability (a potential score of 5), the final assessment of the knowledge presented in both versions of the B&S course is a score somewhere between a 3 and a 4. Despite the broader applicability of the 'skill-based knowledge,' course content is strongly aligned to biological themes. The presentation of knowledge tends not to venture in other areas of science or even to psychology or social theory (focussing mainly on media and public perspectives) – which ultimately means that there are still elements of knowledge which are "treated as unique to a subject area" (The School Reform Longitudinal Study Research Team, 2001, p.3).

There were variations for the score assigned to the dimension 'problematic knowledge', but differences seemed to relate more to session type (guest lecture, CC lecture, discussion, and documentary) than course version (Table 5.1). The exception to this was the documentary sessions, where there was an increase in the expectations around the construction of knowledge in the B&S-V2 (a consequence of the CC pausing the documentary at numerous places to ask the students reflective questions).

In both B&S-V1 and B&S-V2, lectures by the CC generally had a problematic knowledge score of 3, whereas guest lectures (in both B&S-V1 and B&S-V2) scored 1-2 depending on the presenter (Table 5.1). In the case of the guest lectures, the score is a product of the fact that most presenters did not ask questions (and those that did, used a low-level cognitive prompt rather than getting students to challenge the construction of knowledge). Whereas the CC lectures did include questions which required students to think about the information sources, alternative perspectives and how to choose between multiple interpretations.

Discussion sessions scored the highest in terms of generating understanding that knowledge production is subject to "political, social and cultural influences" (The School Reform Longitudinal Study Research Team, 2001, p.4). The expectation placed on students by the CC is at a score of 5 ("present all knowledge as problematic"), particularly in the B&S-V2 course where there was more expectation placed on the quality of evidence and argument. Though in practice most students functioned at a score of 4 ("multiple interpretations and constructions of information having equal status"). However, the descriptor for the 4th parameter was not a particularly good fit for this course, as there is the expectation that peer-reviewed scientific evidence should be given more weight than opinion (i.e. "constructions of information presented as having equal status") as the score 4 descriptor states (The School Reform Longitudinal Study Research Team, 2001, p.4).

Table 5.1 Comparison of scores for the QSRLS dimension problematic knowledge for the two versions of the biology and society course.

Session type	Score	
	B&S-V1	B&S-V2
Guest Lecture	1 or 2	1 or 2
Topic Coordinator Lecture	3	3
Discussion	4 or 5	4 or 5
Documentary	1	2

For the final QSRLS dimension used in this study, the courses were evaluated by the dimension problem-based curriculum. There was only one difference between the two versions of the course (Table 5.2). This occurred during the discussion sessions from the B&S-V2 course, where some of the questions and problems posed to the students would occasionally occupy a large portion of class time. In both versions of the course, discussion sessions scored the highest on the dimension scale, with guest lectures and documentaries providing the fewest opportunities for students to engage their problem-solving skills (Table 5.2).

Table 5.2 Comparison of scores for the QSRLS dimension problem-based curriculum for the two versions of the biology and society course.

Session type	Score	
	B&S-V1	B&S-V2
Guest Lecture	1	1
Topic Coordinator Lecture	2	2
Discussion	3	3 occasional 4
Documentary	1	1

Observation tools findings: EQUIP

The EQUIP tool was used to reflect on classroom experiences but some extra codes concerning the role of the student and instructor were considered to better classify the types of learning experiences in two versions of the B&S course.

Trends across the EQUIP indicators (including discourse[16] and instructor[17] factors) revealed that the B&S-V1 course generally functioned at a pre-inquiry level,[18] despite the fact that regular opportunities were provided to work at higher inquiry levels (Table 5.3). This is largely a consequence of 53% of the course being delivered by guest lecturers (refer back to Figure 5.2), where only one presented in a similar manner to the CC (refer back to Figure 5.7A). If the entire course was delivered like the CC lectures and discussion sessions, the overall inquiry level would have been a level 2 (developing) or 3 (proficient). While the guest lecturer's presentations introduced students to key issues and theory, it was done in a very passive manner where the student's role was to receive the information. In contrast, the CC's instructional strategies shift from educator-centred to student-centred (facilitating students' engagement in thinking and reflection activities and creating a classroom environment where students are active learners) (Table 5.3). Similarly, evaluation of the CC's questioning style against the EQUIP discourse factors also indicates that these sessions placed a greater expectation on students to draw a conclusion and justify their response.

The main differences between B&S-V1 and B&S-V2 related to the lecture and discussion components, where a great range of inquiry related activities and higher cognitive levels were included in B&S-V2 (Table 5.3). This is especially evident in the shift from developing to a proficient inquiry in activities led by the CC (for example, lecture and discussion sessions).

---

[16] The EQUIP protocol Discourse Factors measure "classroom climate and interactions relating to inquiry instruction" (Marshall et al., 2009, p.51). The factors used in this study include the include question level, question complexity, and classroom interactions.

[17] The constructs associated with instructional factors centre the role of the instructor and student in the classroom (Marshall et al., 2009).

[18] The EQUIP creators designate four levels for inquiry - pre-inquiry, developing inquiry, proficient inquiry and exemplary inquiry. However, the descriptor for each inquiry level is slightly different for each construct being measured. See Appendix E: Table E.3 which is adapted from in Marshall et al., 2009.

The other aspect of the EQUIP tool used to evaluate the course was the cognitive demand placed on students (See Appendix E: Table E.3 for a list of codes and their descriptions). In the B&S-V1 course five of the six cognitive demand levels were witnessed by the researcher; with the majority of time being spent in the "receipt of knowledge" and "lower order" cognitive stages. In the B&S-V2 course, the cognitive demand was slightly increased. All six cognitive levels were observed (though only five regularly), with sessions delivered by the CC providing more opportunities to engage in the ‘apply’ and ‘analyse/evaluate’ thought modes. Discussion sessions consistently placed the highest cognitive demand on students, particularly in B&S-V2.

Table 5.3 Summary of the EQUIP results for the two versions of the biology and society course. The green boxes represent the trends displayed in B&S-V1, and the orange boxes represent the trends in B&S-V2. Definitions for each of these components can be found in Appendix E: Table E.3 and E.4.

EQUIP component		Level	Levels displayed							
			Guest Lecture		Lecture		Discussion		Documentary	
		Other	B&S-	V2*	V1*	V2*	V1	V2*	V1	V2*
Cognitive Level		Receipt of Knowledge								
		Lower Order	*	*						
		Apply								
		Analyze/evaluate								*
		Create or transfer						*		
Instructional factors	Teacher Role	Preinquiry								
		Developing Inquiry								
		Proficient Inquiry								
		Exemplary Inquiry								
	Student Role	Preinquiry								
		Developing Inquiry								
		Proficient Inquiry								
		Exemplary Inquiry						*		
Discourse factors ~	Questioning Level	Preinquiry								
		Developing Inquiry								
		Proficient Inquiry								
		Exemplary Inquiry						*		
	Complexity of Questions	Preinquiry								
		Developing Inquiry								
		Proficient Inquiry								
		Exemplary Inquiry								
	Classroom Interactions	Preinquiry								
		Developing Inquiry								
		Proficient Inquiry								
		Exemplary Inquiry						*		

*observed but not regularly

~most common level achieved displayed

Student perceptions of classroom experiences

The majority of students thought they got to regularly practice most of the 23 learning opportunities mentioned in the perception survey (learning opportunities identified as Q1-23; Table 5.4 A and B). This was true of both B&S-V1 and B&S-V2. The two skills that both cohorts of students thought they did not get to practice regularly involved math (Q12) and data collection (Q22).

There were also differences between the two cohort's perceptions of the skills they thought the course allowed them to practice. For example, a higher proportion of students from the B&S-V1 cohort thought they got regular practice at evaluating spurious relationships (Q5), identifying suitable solutions (Q13), and explaining the best solution (Q14). Interestingly, this cohort actually had fewer opportunities to engage with discussions and problem-solving tasks than the B&S-V2 cohort (refer back to Figure 5.3A and 5.3B).

A higher proportion of students from the B&S-V2 cohort indicated they thought the course provided opportunities to practice information literacy skills (such as summarising patterns/drawing conclusions, see Q1, Q8), identifying additional information needed (see Q4/Q7), providing alternative interpretations (see Q9) and explanations (see Q3), and understanding how changes to a situation might affect the solution (see Q15) than in B&S-V1. Additionally, a higher proportion of students from B&S-V2 thought they got regular practice at 6 of the 8 non-CT related skills than in the B&S-V1 cohort (Table 5.4B). However, the most interesting result from B&S-V2 perceptions is that slightly fewer (8% less) students in B&S-V2 thought they got to practice solving real-world problems. This is despite the fact that the intervention tasks added to B&S-V2 actually increased the number of opportunities to work on these kinds of tasks compared to the experiences offered in B&S-V1.

Table 5.4 (A) Summary of student perceptions of the CT related skills they thought they got to practice in the two versions of the biology and society course.

Survey descriptor: CAT measurable skills		Perceptions of learning opportunities					
		practised regularly		practised irregularly		never practised	
		B&S-V1	B&S-V2	B&S-V1	B&S-V2	B&S-V1	B&S-V2
Q1	summarize a pattern of information without making inappropriate inferences.	41%	71%	45%	26%	14%	3%
Q2	evaluate how strongly correlational-type data supports a hypothesis.^	52%	53%	41%	47%	7%	0%
Q3	provide alternative explanations for observations.^	66%	83%	31%	17%	3%	0%
Q4 / Q7	identify additional information needed to evaluate a hypothesis or particular explanation of an observation.^	32%	69%	61%	31%	7%	0%
Q5	evaluate whether spurious relationships strongly support a claim.^#	69%	49%	28%	51%	3%	0%
Q6	provide alternative explanations for spurious relationships.	66%	56%	31%	44%	3%	0%
Q8	determine whether an inference in an advertisement is supported by information.	62%	66%	34%	31%	3%	3%
Q9	provide relevant alternative interpretations of information.^	68%	77%	25%	23%	7%	0%
Q10	separate relevant from irrelevant information when solving a real-world problem.^#	69%	82%	28%	18%	3%	0%
Q11	analyse and integrate information from separate sources to solve a real-world problem.^#	86%	82%	10%	15%	3%	3%
Q12	use basic mathematical skills to help solve a real-world problem.	7%	17%	62%	51%	31%	31%
Q13	identify suitable solutions for a real-world problem using relevant information.^#	90%	82%	10%	18%	0%	0%
Q14	identify and explain the best solution for a real-world problem using relevant information.	93%	80%	7%	20%	0%	0%
Q15	explain how changes in a real-world problem situation might affect the solution.^#	86%	89%	14%	11%	0%	0%

^ denotes skill potentially affected by pedagogical changes in version 2 of the B&S course.

# denotes skill targeted by an intervention task.

Table 5.4 (B) Summary of student perceptions of the other skills they thought they got to practice in both versions of the B&S course.

	Survey descriptor: other course skills	Perceptions of learning opportunities					
		practised regularly		practised irregularly		never practised	
		B&S-V1	B&S-V2	B&S-V1	B&S-V2	B&S-V1	B&S-V2
Q16	explain the methods of science.	100%	74%	0%	26%	0%	0%
Q17	consider why scientific knowledge is testable.	100%	79%	0%	21%	0%	0%
Q18	explain why scientific knowledge is testable by further inquiry.	62%	77%	38%	23%	0%	0%
Q19	explain the role of science in society.	86%	94%	14%	6%	0%	0%
Q20	explain the relevance of science in society.	72%	94%	28%	6%	0%	0%
Q21	design and plan an investigation	28%	40%	55%	54%	17%	6%
Q22	collect and accurately record scientific data.	17%	35%	55%	44%	28%	21%
Q23	interpret and draw conclusions from scientific data.	38%	83%	52%	11%	10%	6%

Student perceptions of their CT and general science abilities

Student perceptions of their CT and general science abilities were explored using different items between B&S-V1 and B&S-V2 (as outlined in the methods, see Section 5.2.2). This section is one of two places where it is not possible to do a direct comparison between the B&S-V1 and B&S-V2 courses.

Students in B&S-V1 completed a set of questions that asked them to reflect on skills they thought the course had helped them develop (as opposed to indicating their belief about their current skill level). Students agreed with 19 of 23 the statements indicating they thought the B&S course helped them develop these skills (Table 5.5). The pattern of responses to these items mirrored the responses to the 'practice' survey items (Table 5.4A and 5.4B), with students indicating they disagreed that the course had helped them develop 'math skills' (Q12) and 'experimental design and data collection skills' (Q22).

In B&S-V2, students were asked to reflect on their current CT abilities (as well as some more general science skills) at the start and end of the semester. There were statistically significant changes in student perception of their abilities at the start and end for all questions except Q12 and Q22 (See Table 5.6 and 5.7 respectively). Interestingly, this pattern of pre-test post-test results is consistent with student perceptions of the skills they thought the course developed in B&S-V1 (where only a post-test was used).

Table 5.5 Summary of the B&S-V1 perception survey results for the CT and general science skills and abilities students thought the course helped them development.

	Survey descriptor	Perception		
		agree	disagree	unsure
	CAT measurable skill/abilities			
Q1	Summarize a pattern of information.	64%	36%	0%
Q2	Evaluate how strongly data supports a hypothesis.	73%	19%	8%
Q3	Provide alternative explanations for observations.	93%	7%	0%
Q4 / Q7	Identify additional information needed to evaluate a hypothesis or claim.	71%	29%	0%
Q5	Evaluate contentious claims.	89%	11%	0%
Q6	Provide alternative explanations for observations.	93%	7%	0%
Q8	Determine whether an inference is supported by information.	82%	18%	0%
Q9	Provide alternative interpretations of information.	93%	7%	0%
Q10	Separate relevant from irrelevant information when solving a real world problem.	79%	21%	0%
Q11	Integrate information from separate sources to solve a real world problem.	93%	7%	0%
Q12	Use basic mathematical skills.	14%	68%	18%
Q13	Identify suitable solutions for a real world problem.	92%	8%	0%
Q14	Explain the best solution for a real world problem using relevant information.	93%	7%	0%
Q15	Explain how changes in a real world problem situation might affect the solution.	85%	15%	0%
	General science skill/abilities			
Q16	Explain the role of science in society.	96%	4%	0%
Q17	Explain the relevance of science in society.	100%	0%	0%
Q18	Explain the methods of science.	64%	36%	0%
Q19	Understand why current scientific knowledge is contestable.	85%	15%	0%
Q20	Understand why scientific knowledge is testable.	79%	21%	0%
Q21	Understand how to design and plan an investigation.	39%	50%	11%
Q22	Understand how to collect and accurately record scientific data.	21%	71%	7%
Q23	Understand how to interpret and draw conclusions from scientific data.	36%	64%	0%

Table 5.6 Summary of the B&S-V2 cohort perception survey results for questions relating to CAT measurable skills. Significance explored using a Wilcoxon signed-rank test statistic (*W*), with a two-sided p-value. In all cases, alpha was set to 0.05, where bolded p-values indicate where a significant difference lies and the asterisk denote the level of significance, * <0.05, ** <0.01, * <0.001.**

	Q1		Q2		Q3		Q4		Q5	
	pre	post	pre	post	pre	post	pre	post	pre	post
Maximum	5	5	5	5	5	5	5	5	4	5
Median	3	4	4	4	4	4	4	4	3	4
Mean	3.44	4.18	3.46	4.09	3.56	4.06	3.63	4.14	2.97	3.97
Statistical test statistic	**195**		**256**		**125**		**154**		**389**	
Statistical test: p-value	***		***		*		***		***	

	Q6		Q7		Q8		Q9		Q10	
	pre	post	pre	post	pre	post	pre	post	pre	post
Maximum	4	5	5	5	4	5	4	5	5	5
Median	3	4	4	4	4	4	4	4	4	4
Mean	2.74	3.94	3.63	4.14	3.41	4.06	3.51	4.23	3.80	4.23
Statistical test statistic	**435**		**154**		**213**		**260**		**156**	
Statistical test: p-value	***		***		***		***		*	

	Q11		Q12		Q13		Q14		Q15	
	pre	post	pre	post	pre	post	pre	post	pre	post
Maximum	5	5	5	5	5	5	5	5	5	5
Median	4	4	4	4	4	4	4	4	4	4
Mean	3.60	4.09	3.76	3.76	3.86	4.15	3.74	4.09	3.94	4.31
Statistical test statistic	**154**		-8		**56**		**79**		**104**	
Statistical test: p-value	***		ns		*		**		**	

Table 5.7 Summary of the B&S-V2 cohort perception survey results for questions relating to general science skills and abilities. Significance explored using a Wilcoxon signed-rank test statistic (W), with a two-sided p-value. In all cases, alpha was set to 0.05, where bolded p-values indicate where a significant difference lies and the asterisk denote the level of significance, * <0.05, ** <0.01, * <0.001.**

	Q16		Q17		Q18		Q19	
	pre	post	pre	post	pre	post	pre	post
Maximum	5	5	5	5	5	5	5	5
Median	4	4	4	4	4	4	4	4
Mean	3.62	4.06	3.54	4.20	3.71	4.14	3.91	4.43
Statistical test statistic	**150**		**197**		**81**		**136**	
Statistical test: p-value	**		***		**		***	

	Q20		Q21		Q22		Q23	
	pre	post	pre	post	pre	post	pre	post
Maximum	5	5	5	5	5	5	5	5
Median	4	4	4	4	4	4	4	4
Mean	4.00	4.34	3.51	3.94	3.83	3.97	3.80	4.18
Statistical test statistic	**78**		**125**		35		**78**	
Statistical test: p-value	**		*		ns		**	

Student perceptions of their disposition to transfer their knowledge and skills

When considering student perceptions of their disposition to think critically there were 4 disposition items in common between B&S-V1 and B&S-V2 (Table 5.8), and 11 items that were unique to B&S-V2 (Table 5.9). For the items that were common between the two course versions, the same proportions of students in each B&S cohort indicated they use a critical approach to analysing data and arguments in their daily lives. There were a few differences between the post-test perspectives of the two cohorts. A higher proportion of students in the B&S-V1 cohort indicated they believed they connect and transfer ideas they learn in class to other situations (Q4 and Q25). However, for Q26, a higher proportion of students in the B&S-V2 cohort indicated they used systematic reason to approach problems. Overall, at least 76% of students in both cohorts agreed with each of the four post-test disposition statements (Table 5.8).

Table 5.8 Summary of post-test responses to the four common disposition-related perspective items given to students in B&S-V1 and B&S-V2.[19]

	Statement	agree		disagree		unsure	
		B&S-V1	B&S-V2	B&S-V1	B&S-V2	B&S-V1	B&S-V2
Q24	I am in the habit of connecting key ideas I learn in my classes with other knowledge.	93%	83%	7%	17%	0%	0%
Q25	I am in the habit of applying what I learn in classes to other situations.	90%	83%	10%	14%	0%	3%
Q26	I am in the habit of using systematic reasoning in my approach to problems.	76%	85%	24%	15%	0%	0%
Q27	I am in the habit of using a critical approach to analysing data and arguments in my daily life.	76%	76%	24%	24%	0%	0%

For the disposition items that only the B&S-V2 cohort completed, at least 59% of the cohort agreed with the 14 out of the 15 disposition statements at the pre-test. However, there was still a positive shift in agreement of at least 10-38% for seven of the disposition statements in the post-test (Table 5.9: Q24, Q25, Q26, Q27, Q28, Q32, Q33). Three of these shifts were statistically significant (Q24, Q26 and Q32). There were also four statements where post-test proportions were lower (Table 5.9: Q30, Q34, Q35, Q37), though not significantly so. There was one other statistically significant change in student's assessment of their dispositions (Q38). This result is interesting because this shift in perception is not evident in the percentage values presented (where the 5-point Likert scale was reduced to 3 categories); however, it is visible in the plot of differences which factors in the trends arising from matched-pair data (see Appendix K: Figure K.6).

[19] Not every student who completed the pre and post CAT test completed all survey items therefore B&S-V1 (n= between 28-29– some students skipped an item) and B&S-V2 (n= between 33-35 – some students skipped an item).

Table 5.9 Summary of the B&S-V2 cohort responses for the disposition questions on the student perception survey (n= between 32-35 – some students skipped a few items). Significance explored using a Wilcoxon signed-rank test statistic (*W*), with a two-sided p-value. Alpha was set to 0.05, where bolded p-values indicate where a significant difference lies, asterisk denote the level of significance, * <0.05, ** <0.01, * <0.001.**

	Statement	agree		disagree		unsure		significance
		pre	post	pre	post	pre	post	
Q24	**I am in the habit of connecting key ideas I learn in my classes with other knowledge.**	**61%**	**83%**	**39%**	**17%**	**0%**	**0%**	**
Q25	I am in the habit of applying what I learn in classes to other situations.	73%	83%	27%	14%	0%	3%	
Q26	**I am in the habit of using systematic reasoning in my approach to problems.**	**47%**	**85%**	**53%**	**15%**	**0%**	**0%**	**
Q27	I use a critical approach to analyzing data and arguments in my daily life.	59%	76%	41%	24%	0%	0%	
Q28	I usually try to think about the bigger picture during a discussion.	73%	86%	27%	14%	0%	0%	
Q29	I often use new ideas to shape (modify) the way I do things.	76%	82%	24%	18%	0%	0%	
Q30	I often re-evaluate my experiences so that I can learn from them.	82%	77%	18%	23%	0%	0%	
Q31	I use more than one source to find out information for myself.	94%	97%	6%	3%	0%	0%	
Q32	**I am often on the lookout for new ideas.**	**64%**	**83%**	**36%**	**17%**	**0%**	**0%**	*
Q33	I usually check the credibility of the source of information before making judgements.	70%	83%	30%	17%	0%	0%	
Q34	I sometimes find a good argument that challenges some of my firmly held beliefs.	73%	68%	27%	29%	0%	3%	
Q35	It is important to understand other people's viewpoint on an issue.	97%	94%	3%	6%	0%	0%	
Q36	I usually think about the wider implications of a decision before taking action.	79%	82%	21%	18%	0%	0%	
Q37	It is important to justify the choices I make.	85%	82%	15%	18%	0%	0%	
Q38	**I often think about my actions to see whether I could improve them.**	**79%**	**79%**	**21%**	**21%**	**0%**	**0%**	*

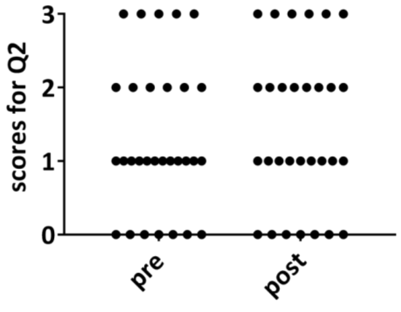
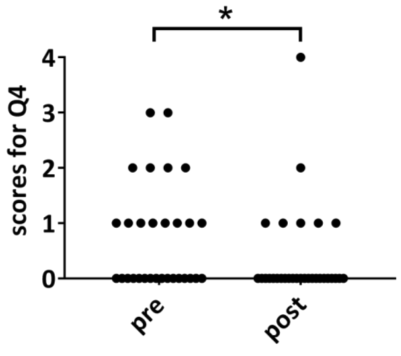
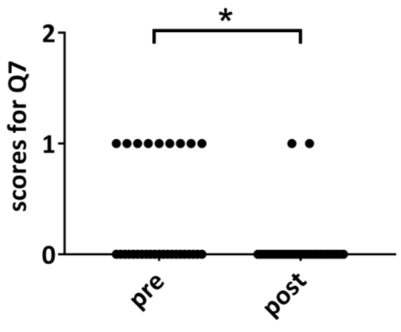
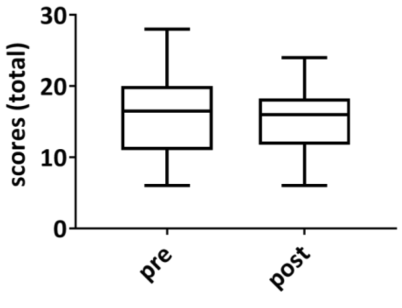
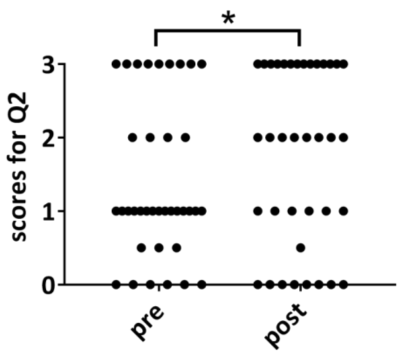
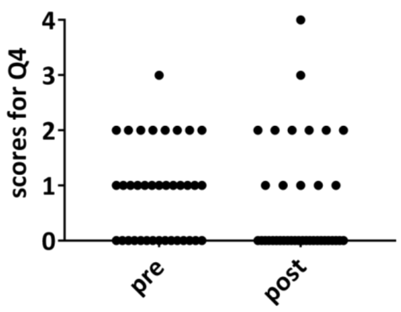
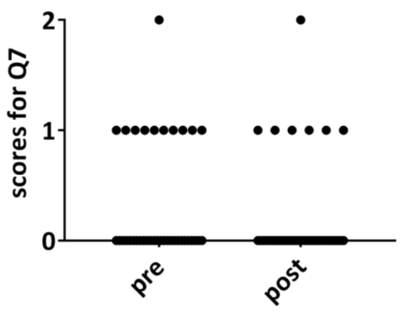
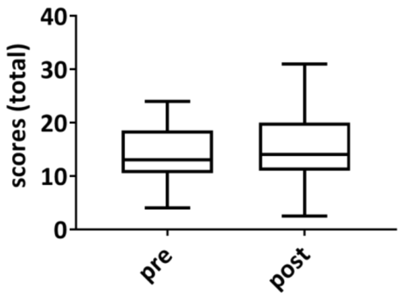
5.4.2 Summary of CT performance results

In B&S-V1, there was a non-significant decrease in the overall CAT performance across the semester. At a question level, there were five questions with a non-significant decrease (Q3, Q6, Q9, Q11 and Q15 - see Appendix K Table K.1), and two questions with a significant decrease in student performance (Q4: $W = -57$, p-value = 0.048 and Q7: $W = -35$, p-value = 0.039). Interestingly, both of these questions related to the same skill - "identifying additional information needed to evaluate a hypothesis" (Center for Assessment and Improvement of Learning, 2013, p.24). The maximum score achieved on the post-test was lower than the pre-test max, and the range of overall scores achieved on the post-test was also smaller (Table 5.9; a full summary of the performance results can be found in Appendix K: Table K.1). Exploration of the mean scores obtained by the B&S-V1 cohort were used to identify opportunities to improve student CT performance. For example, there were only 6 out of the 15 questions where, on average, more than 50% of the available points were scored (Q1, Q5, Q8, Q10, Q12, and Q14 - Appendix K: Figure K.3 and Table K.1). Comparing the difference between the total available points (38) and the mean post-test score (post-test mean = 15.20 ± 4.53) also revealed that most students only scored around 40% of the total available points. Intervention tasks were modelled on those questions with CT skills that were relevant to course learning outcomes where there was little to no improvement (such as Q3, Q6, Q7, Q9, Q11, and Q15).

The trends in the B&S-V2 cohort scores were more indicative of improvement in students CT skills resulting from the course. In B&S-V2, the overall pre-test and post-test scores increased by 1.2 points and the interquartile range is greater (Table 5.10). In addition, the increase in performance on Q2 was determined to be statistically significant (Appendix K: Table K.2). However, like the B&S-V1 cohort, performance on Q4 and Q7 still declined, but not significantly. Examination of mean scores across the questions revealed there was still room for further skill improvement (and therefore pedagogical changes). Only 7 out of the 15 questions had mean scores that were more than 50% of the available points (Q1, Q2, Q5, Q6, Q8, Q10, and Q12), and again most students only scored around 40% of the total available points (Appendix K: Figure K.4). However, these apparent low scores and general lack of questions with statistically significant improvements do not negate the fact that overall, the B&S-V2 course led to stronger CT performance outcomes.

Comparison between the two cohort's performance changes further demonstrate the positive influence of the B&S-V2 course on student CT performance. As Figure 5.9 shows, the performance by the B&S-V2 cohort was greater on 13 out of 15 questions and the magnitude of the increase was at least 5% greater for 8 of these questions. As shown in Figure 5.9, performance improved on all questions targeted by an intervention (working from the assumption that if no changes were made to the course trends observed in B&S-V1 would have been repeated in B&S-V2). There were two statistically significant changes in CAT performance between the two course versions (Figure 5.10). One of these related to CAT Q6, a skill targeted by the pedagogical changes (U = 410, p-value = 0.043). The other was for the overall CAT performance (U = 385, p-value = 0.031). Importantly, the median score change was higher for B&S-V2 compared to B&S-V1 (B&S-V2 median change = 1; B&S-V1 median change = 0) and upwards of 60% of the B&S-V2 cohort experienced an increase in their CAT score between the pre-test and post-test. Whereas in B&S-V1, the median change was zero, with approximately 33% of the cohort achieving a score change that was greater than zero.

Table 5.10 Summary of the B&S cohort's significant pre-test and post-test CAT results (B&S-V1 n = 30; B&S-V2 n= 37). Due to variations in the scales across questions, different statistical tests were applied as appropriate. These are indicated through the symbols near the question number and are explained below in the table notes. In all cases, alpha was set to 0.05, where bolded p-values indicate where a significant difference lies and the asterisk denote the level of significance, * <0.05, ** <0.01, * <0.001. A complete summary of the CAT results can be found in Appendix K: Table K.1 and K.2.**

Course	CAT Q2^	CAT Q4^	CAT Q7^	Total CAT score#
B&S-V1				
B&S-V2				

^ Wilcoxon signed-rank test statistic (*W*), with a two-sided p-value. A non-parametric version of the paired t-test, to assess whether the matched population mean ranks differ between pre and post.

paired t-test statistic (*t*), with a two-sided p-value. A parametric test after confirmation of normality (using the Shapiro-Wilk and D'Agostino & Pearson normality tests) to assess whether population means differ pre and post.

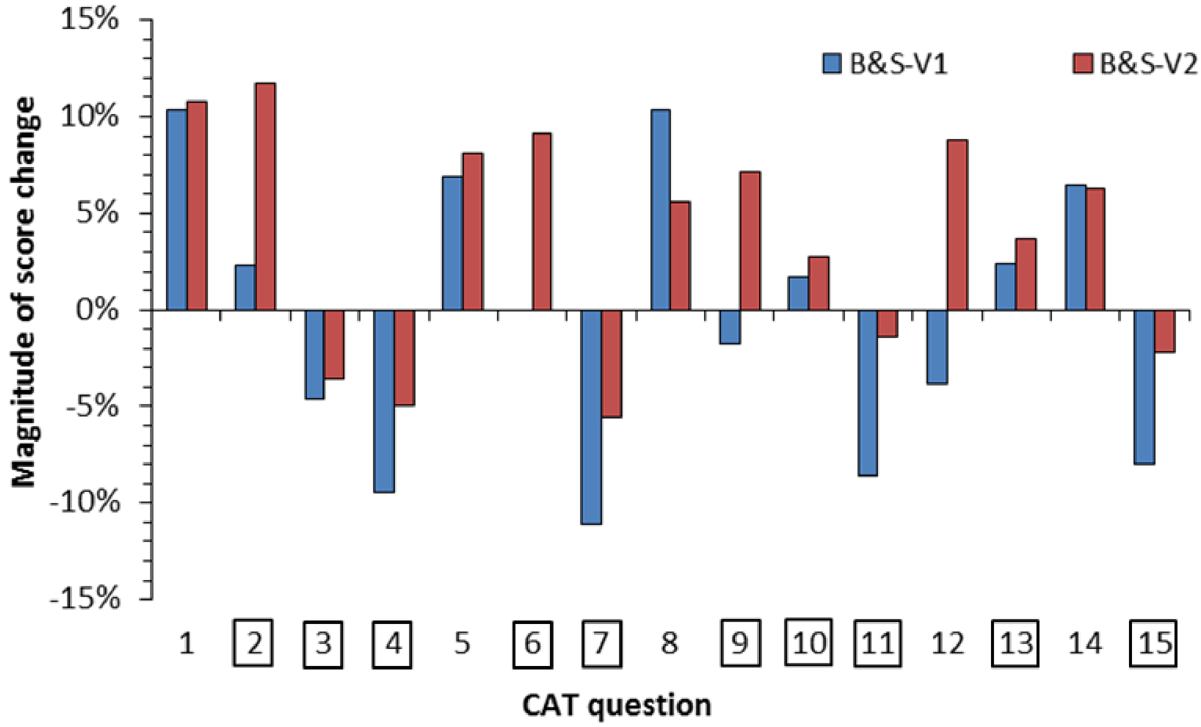


Figure 5.8 Comparison of the B&S-V1 and B&S-V2 course CT performance changes across CAT questions. The boxed question numbers signify questions with skills targeted through one or more of the interventions / pedagogical changes introduced in B&S-V2.

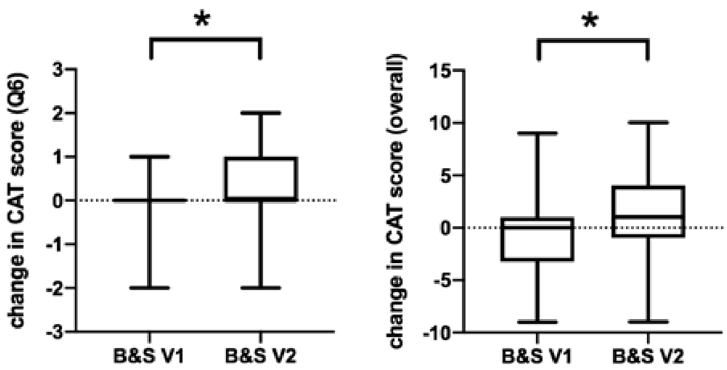


Figure 5.9 Comparison of the change in Q6 and overall CAT performance between the two versions of the B&S course. Significance explored using a Wilcoxon signed-rank test statistic (*W*), with a two-sided p-value where alpha was set to 0.05. The asterisk denotes the level of significance, * <0.05, ** <0.01, * <0.001. A complete summary of the change in CAT performance can be found in Appendix K: Figure K.5.**

Comparison to American students

Comparison of the Australian and American students by the proportion of points achieved across the CAT revealed that overall pre-test and post-test scores for both versions of the B&S course were higher than the freshman norm value (Figure 5.11). However, a one-sample means test using the freshman norm values[20] revealed that only the pre-test performance from the B&S-V1 cohort and B&S-V2 post-test performance were statically significantly higher than the norm values (t = 2.095, p-value = 0.04; t = 2.305, p-value = 0.027, respectively).

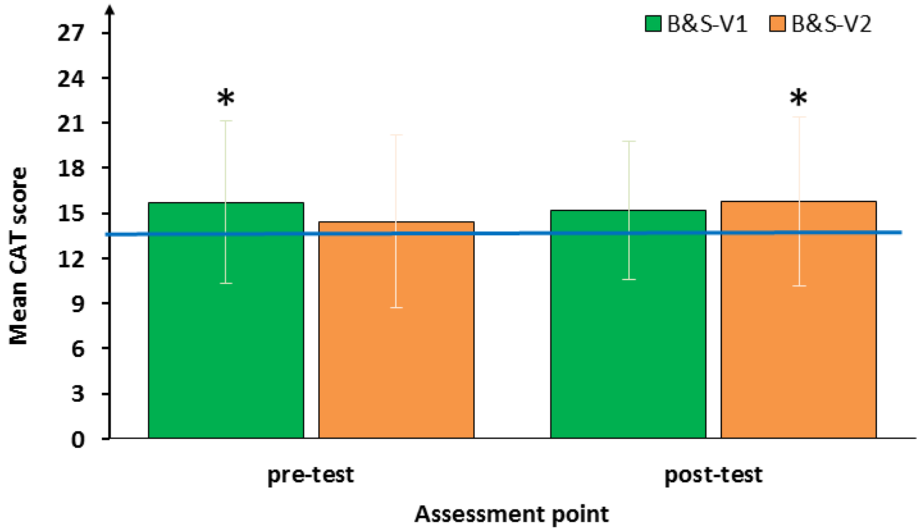


Figure 5.10 Mean pre-test and post-test CAT performance results for version 1 and 2 of the biology and society course. The navy line indicates the typical score of American Freshmen based on information from the CAT creators. The B&S-V1 cohort's pre-test score and the B&S-V2 cohort's post-test score were significantly different from mean freshman score, at a 0.05 significance level (the asterisk denotes the level of significance, * <0.05). The error bars represent +/- 1 SD. The maximum possible test score is 38.

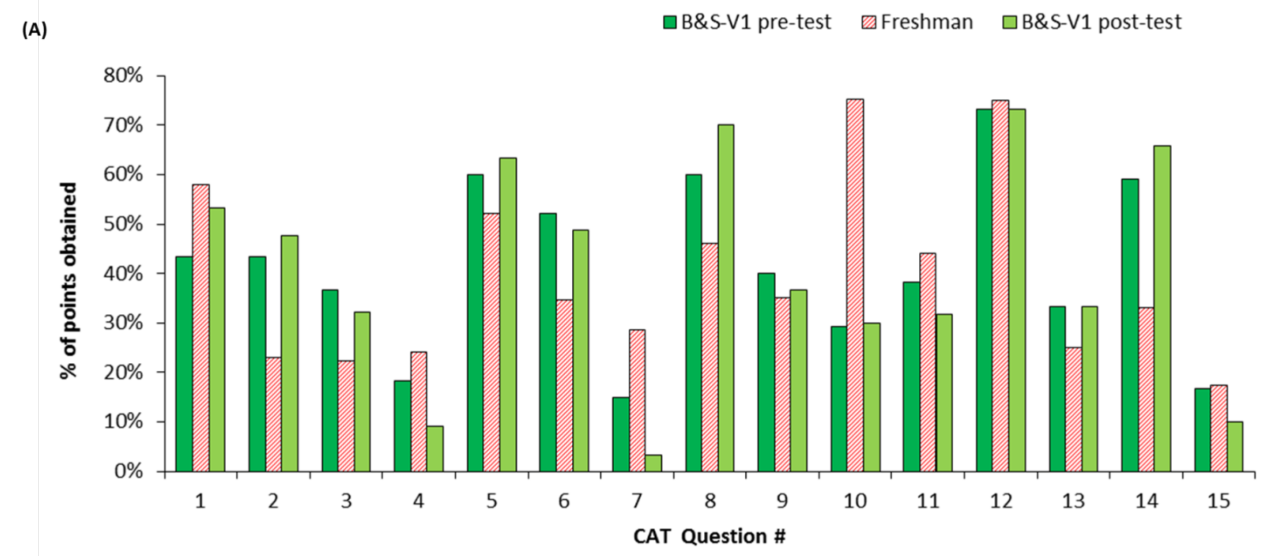
Looking at the B&S-V1 on a question by question level, there were eight questions where the cohort's performance was higher than the freshman (Figure 5.12A: Q2, Q3, Q5, Q6, Q8, Q9, Q13, and Q14). But of these, there were only four questions in the pre-test (and three questions in the post-test) where the B&S-V1 cohort's performance was significantly higher (See Appendix K: Table K.3: Q2, Q3, Q6, and Q14, and Table K.4: Q2, Q6, Q8, and Q14). In terms of the seven questions where the B&S cohort scored lower than the freshman (Figure 5.12A: Q1, Q4, Q7, Q10, Q11, Q12, and Q15), only one of these (Q7) was significantly lower than the norms in the pre-test and two (Q4 and Q7) in the post-test results (Appendix K: Table K.3, p-value <0.05). Interestingly, these two questions portray the same skills set – identifying additional information needed to evaluate a hypothesis.

Patterns for the B&S-V2 cohort compared to the freshman norms were slightly different from the B&S-V1 cohort results. There were five questions (two fewer) where the freshman outperformed the B&S cohort on both the pre-test and post-test (Figure 5.12B). Pre-test and post-

[20] Supplied by the CAT developers in the CAT report (Appendix K: Table K.3)

test performance by the B&S-V2 cohort was statistically significantly than the freshman lower on Q15. For the seven questions where the B&S-V2 cohort performance was higher than the freshman norms (Figure 5.12B: Q2, Q3, Q4, Q5, Q6, Q8, Q13, and Q14), four of these (Q2, Q5, Q6, and Q8) were statistically significantly higher (Appendix K: Table K.5 and Table K.6).

Between B&S-V1 and B&S-V2 there was a consistent trend concerning which questions the Australian performance was lower than the freshman. For example, both B&S cohorts scored less than the freshman on Q4, Q7, Q10, Q11, and Q15.



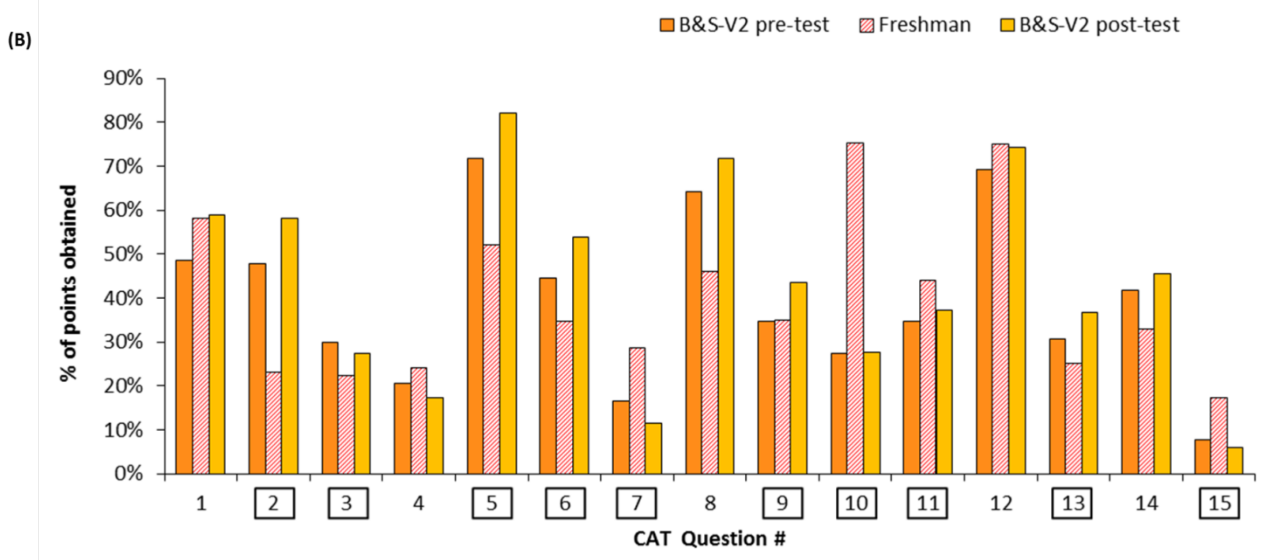


Figure 5.11 A-B Comparison of the two biology and society cohorts against the American freshman norms, based on the percentage of points obtained per CAT question. (A) contains the B&S-V1 pre-test and post-test results, and (B)

the B&S-V2 pre-test and post-test results. Boxes around question numbers in Figure B indicate the questions with skills targeted by B&S-V2 course changes.

Comparisons of student perception of skills, practice and CT performance

Despite trends in student perceptions of their skills levels, which demonstrated a statistically significant positive shift in their assessment of their CT abilities, all but one of the CAT performance results revealed non-significant increases in performance. However, comparison of the overall change in performance revealed a significant increase in CT skills between B&S-V1 and B&S-V2. This suggests the pedagogical changes in B&S-V2 led to stronger performance outcomes. In addition, the fact so many of the skill perceptions and disposition results were significantly positive (and the skills pattern matched the B&S-V1 cohort) suggests that some aspect of this course (aka the explicit discussion of CT) instils confidence in a student's ability and willingness to engage in the CT processes.

Overall, based on Mantel's analysis there were no correlations between the CAT performance, CAT skill perception and CAT skill practice responses. There were also no correlations between CAT performance and course grade, irrespective of pre-test performance or overall change on the CAT in B&S-V1 or B&S-V2.

5.5 Discussion

This chapter has investigated the role of an Australian science and society course, biology and society (B&S), on the development of students' CT skills. Findings from similar courses taught in the USA indicate that science and society courses have a positive effect on CT development (Rowe et al., 2015; Gottesman and Hoskins, 2013), but explorations had not been undertaken in an Australia context. This course choice enabled exploration of the assumption that science and society courses would produce a positive effect on CT development in Australian tertiary science classrooms. This scientific literacy heavy course, which includes critically analysing themes and arguments, also provided a contrast to the pure science course explored in Chapter 4. This enabled a broader understanding of CT development outcomes in science courses in different learning contexts.

This chapter is the second in a series of studies on CT development that focus on generating understanding about Australian tertiary science classrooms (part of aim 2) and modelling a process of assessing CT development (part of aim 3). However, this chapter also extends understanding of the assessment model, by demonstrating how the evaluation process provides sufficient feedback to be able to discern and implement changes to a course that will subsequently improve student CT development. Thus, this study helps address objective one (knowledge generation) and objective two (outcome comparisons) of aim three (refer to Chapter 4, Section 4.2).

5.5.1 Effect of biology and society on CT development

Biology and society is a course designed to train students how to understand and interpret different presentations of scientific ideas, as well as to develop students written and verbal reasoning and communication skills (in other words, scientific literacy skills). Training scientifically literate citizens through school/university is important for equipping students with skills to navigate the increasing volume of fake news (Nagi, 2018), and for fostering skills for mindful engagement in democratic societies (Arum et al., 2016). But perhaps more importantly scientific literacy will help individuals to make sound judgements and be better critical thinkers. Scientific literacy is useful for training individuals how to weigh up opinion versus evidence, and negotiate conflicting scientific evidence (such as issues pertaining to climate change), as well as being wary of personal data misuse (including, but not limited to, recognising potential identify theft). These types of skills, including CT, are especially important for the biology and society students in this study since most of them are science majors who will need these skills to succeed in their profession. So how effective is this course at developing students' CT abilities?

In the B&S-V1 course, students' CT performance on the CAT declined, yet student perceptions of their CT abilities and dispositions indicated they felt the course had helped them to become better critical thinkers. In B&S-V2, there was a significant improvement in student's perception of their CT skills between the start and end of the semester, as well as some improvement in student performance on the CAT. Students in both versions of the course also indicated they thought it gave them regular opportunities to develop abilities associated with CT. This result is particularly interesting because the pattern of student responses between the two versions of the course were the same, despite having been asked about their CT skill development in slightly different ways. For example, in B&S-V1 students completed a post-test where they expressed which skills they thought the course helped them improve; whereas in B&S-V2 students completed a pre-test and post-test asking them about their skills at that point in the semester.

Both versions of the course also asked students about the skills they thought the course gave them the opportunity to practice, and the pattern was the same here as well. This suggests the explicit discussion of CT made the students more aware of their abilities, whereas in chemistry where everything was embedded, students had lower awareness of their CT abilities, even though based on their CAT performance they had improved. This finding provides support for the role of explicit pedagogical techniques and the importance of disposition in the CT development process (see also Abrami et al., 2015; Kuhn, 2019).

The trends of results published in education literature indicate that anecdotal evidence and student perception of improvement is enough to satisfy some educators of a positive learning outcome (Ebert-May et al., 2011; Andrews and Lemon, 2015). However, there are others that prefer to work from empirical evidence. To these educators, quantitative CT performance findings would provide more weight to the claim the course is effective at developing student CT abilities, and increase their likelihood of adopting the recommendations. Surprisingly, Andrews and Lemon (2015) found that despite the belief that science educators would be convinced to change their teaching approaches if they were presented with data, many college science educators will work based on their personal experiences until they feel an approach is no longer effective. Nevertheless, I have continued to present a mix of the qualitative and quantitative evidence, to satisfy the subset of educators who prefer empirical data (myself included) but also to address literature recommendations about CT assessment. For example, Facione (1990) recommends assessing CT holistically - which means considering skills and dispositions, as well as a willingness to deploy them. In this instance, the willingness to deploy CT abilities seemed low (particularly in the B&S-V1 post-test where students wrote less and performed worse than they did on the pre-test). In addition, the lack of statistical significance between overall pre-test and post-test performance at the 0.05 level (the usual gatekeeper for science) within each course version suggests minimal development of student CT abilities over the course of a single semester. However, the inclusion of evidence-informed pedagogical changes in B&S-V2 led to significant improvement in overall performance compared to the B&S-V1 CT skill performance outcomes. This support the idea that there were stronger performance outcomes from the second version of the course.

P-values are just one aspect of statistical significance and they should not be the sole basis for the dismissal of effects (Amrhein et al., 2019). This means there could be a more relevant diagnostic measure to help judge the merit of this B&S course. For example, Wasserstein et al.

(2019) suggest context and prior evidence are important aspects for thoughtfully interpreting statistical results. In this instance, the CAT creators have provided data-derived suggestions about the magnitude of the CT outcomes one might expect from a four-year college experience where no specific pedagogy has been incorporated to improve student CT development (a four-point CAT score increase).

Based on the conservative assumption of a linear increase of 1 CAT point per year,[21] the average increase of 1.2 points observed during the semester-long B&S-V2 course is evidence that the course was effective at improving student CT skills. Further if no changes were made to the course, and B&S-V2 cohort performed similarly to the B&S-V1 students (decreasing in performance), the size of the effect is even greater. However, the magnitude of the change is still smaller than that observed in two previous studies of first-year science and society courses which also used the CAT to assess CT (Gottesman and Hoskins, 2013; Rowe et al., 2015). The score increase from the Gottesman and Hoskins (2013) study found a mean difference of 4.4 from their semester-long first-year course. Rowe et al. (2015) study across 5 semesters found mean changes between 3.31 and 8.99 points. These two studies demonstrate that it is possible to achieve a greater degree of change in a first-year science and society course than what was observed in my research. However, there are a few reasons for the difference in magnitude observed in the B&S course compared to these US courses.

Firstly, there are cohort differences to consider. For example, students who study science and society courses in the US are generally not science majors (K. Burke da Silva, personal communication, 1st August 2019). The non-science majors' backgrounds and levels of literacy are more diverse (and lower) than science majors (Cotner, Thompson and Wright, 2017), therefore the potential to remedy these student thinking skills are greater. For example, the non-science majors in the Gottesman and Hoskins (2013) study had lower CT ability than the Australian students even after they completed their scientific thinking course (performance change from a mean of 9.6 to 13), supporting the idea that there are cohort difference between the cohorts who study science and society course in Australia compared to the US. The B&S-V2 post-test course results were also significantly higher than the freshman norm data set, further suggesting a higher ability in the Australian students.

[21] *Reminder:* this is an underestimation as the later years of college tend to produce higher CT development (Arum and Roksa, 2014; Roohr et al., 2017).

The second reason for the magnitude differences is motivation. In the Rowe et al. (2015) study students were incentivised to complete the CAT (received extra credit or with the test actually forming part of their course grade). In fact, the biggest increases in CAT performance in the Rowe et al. (2015) study were associated with the semesters that student CAT performance formed a part of their course grade. The role of motivation also provides some insight into the results from B&S-V1. A study by Bensley et al. (2016) suggested students with low-test taking motivation do worse on low-stakes tests. When scoring the CAT tests, it was evident the students had written shorter responses on their post-tests, suggesting reduced effort and lower motivation. This reduced motivation is not surprising given the timing of the post-test assessment (week 12 when every course they take has assignments due), especially since there were no incentives for these students to perform well on the test. In fact, in B&S-V1 only half the number of students who were present at the CAT pre-test attended the post-test session. However, it does make it more challenging to determine the effect of CT development using an instrument that requires students to provide a detailed written response. Further, Simper, Frank, Kaupp, Mulligan, and Scott (2019) suggest that the "reliability and validity" of CT assessment data "is called into question by the impact of student effort." But given the lack of incentives in this biology and society course, in contrast to the findings from Rowe et al. (2015) study, it is likely the emerging trends are underestimates of the student's actual CT ability.

Course delivery modes are an important factor in CT development (Heijltjes, Van Gog and Paas, 2014; Lumpkin et al., 2015; Styers et al., 2018). Over the two years of data collection, the B&S course was observed to use a combination of explicit and embedded CT activities. This course engages students in modern assessment activities including a Massive Open Online Course (MOOC) and a video-based scientific communication assignment, as well as multiple written expositions about current scientific perspectives. The course exposes students to different opinions and expertise through the uses of guest presenters and incorporates some student-centred learning approaches, such as getting to participate in small group and whole-class discussion during workshop sessions. However, the proportion of passive to active learning time corresponded to a more traditional lecture environment than a student-centred one. These types of passive environments are becoming increasingly associated with poorer student learning outcomes (Detlor et al., 2012; Stevens et al., 2019). This could also explain why CT development results were less than expected based on previous findings from science and society courses. This idea is supported by the findings from the B&S-V2 course results where a 12% increase in active learning time, resulted in stronger improvement to CT outcomes. This finding highlights the importance of an ALE with multiple CT focussed tasks for fostering student CT development. Further benefits for shifting pedagogy toward a more ALE include reducing academic performance gaps among biology students (Haak, HilleRisLambers, Pitre and Freeman, 2011; Eddy and Hogan, 2014; Wright, Eddy, Wenderoth, Abshire, Blankenbiller and Brownell, 2016).

An additional factor to consider about the classroom environment is how CT was presented and what CT development opportunities were on offer. The use of explicit and embedded CT approaches are consistent with both literature recommendations about CT development (Abrami et al., 2015), and suggestions from Australian education policy (Jones et al., 2011). However, the classroom observation data revealed that even though CT was discussed explicitly in the course, opportunities to individually practice CT skills was proportionally lower to time spent listening. Heijltjes et al. (2014) have previously found that the absence of either explicit teaching or CT skill practice in a classroom led to comparable outcomes to having no practice or no explicit instruction. Therefore, it is necessary to have both explicit teaching and opportunities to practice CT skills to produce CT development outcomes. So aside from issues relating to test-taking motivation, and the ratio of passive to active learning time, another key contributing factor was that there were not many opportunities for individual students to practice CT skills, despite best intentions to incorporate CT in the course. In fact, one of the main differences between B&S-V1 and B&S-V2 were the number of opportunities to practice CT-related abilities, and in B&S-V2 where the amount of CT skill practice and explicit instruction was increased, students CT performance improved.

There are also two more general lessons about CT which emerge from these case study results. The first concerns CT assessment. This multi-tool approach provides the diagnostic ability to strategically target of CT skills and identify specific classroom environment changes needed in future versions of the course. The power of this approach was that the combination of tools was collectively able to provide more understanding of the course than would have been identified using the tools individually. For example, the results from the CAT alone would not have been able to explain what particular aspects about the course were contributing to the CT gains; and on its own could not specifically pinpoint how the learning environment might be changed to improve outcomes. The various classroom observations measures were able to characterise the environment, but were not able to assess the learning gains. This is why a multi-tool assessment approach is so important for generating usable assessment data.

The second key finding follows on from this. The collective approach highlighted areas of strengths and weaknesses in student performance, as well as the classroom environment, such that strategic changes could be made. Specifically, the number of documentaries was reduced and three short intervention activities were added. The course coordinators presentation style also changed a bit, mostly in the discussion sessions. These changes resulted in a 12% reduction in passive learning experiences. The overall course structure (2x lectures, 1x workshop per week) did not change, nor did the involvement of guest lecturers. In total, only 25% of teaching sessions were affected by the instructional changes, and this was enough to produce a comparative 1-2-year improvement in student CT learning gains compared to the outcomes from the initial course. This highlights that entire course redevelopment may not be required to achieve better outcomes. In addition, the CC has also noted that the changes were not that hard to make or maintain (N. Hunter, personal communication, 23 July 2019). This CC's experience with this process is a promising outcome considering it is the opposite of trends from Andrews and Lemon (2015) and others (Shell, 2001; Black 2009; Stedman and Adams, 2012) where resistance to long-term change was found - largely driven by the fact that educators are time poor and generally not trained to teach for CT.

5.5.2 Practice Implications

There were some unexpected findings relating to the classroom environment in this B&S course. Given the emphasis on argument, evidence and reflection, there was an implicit expectation that the course would develop student CT skills – yet students CT skills did not improve that much (particularly in the B&S-V1 course). Closer examination of the classroom environment revealed that most of the course involved students listening to instructors rather than engaging in more student-centred tasks. In addition, the way content was presented varied between guest lecturers and the course coordinator (CC). So even though the CC's lectures and discussion sessions had elements of student-centred activities, the overall sum of experiences were passive and did not give students many opportunities to deploy the CT-related abilities being learnt. This discrepancy between the CC's view of the course and the actual classroom environment is not overly surprising. A number of previous studies have found a disconnection between educators' perceptions of practice and the actual classroom experience (Fung and Chow, 2002, Ebert-May et al., 2011). However, the difficulty with making changes to the learning environment in this particular course context is that much of the course is delivered by guest lecturers.

The inclusion of guest lecturers in the B&S course was intentional; however, it resulted in a trade-off where approximately half the course presentation was outside of the CC control. The CC felt that it was important for the students to hear from a variety of expert perspectives, who could give experience-based insight into their experiences of science in society, or their views of science. However, not all of these guests are academics, nor skilled in teaching/presenting. This was evident in the difference in the nature of knowledge treatment – the CC would treat knowledge as something to be challenged, whereas the guest speakers would generally present knowledge as fact. Quite often the guest lecturers were presenting their work or their experiences; whereas the CC would describe the different perspectives in science and society alongside each other, getting students to reflect on the difference in perspective as well as their own stance on an issue. The questions asked by guest lecturers were generally low-level prompts (any questions, do you understand etc.?) whereas the questions from the CC required higher cognitive demand – asking for a justified opinion, scientific evidence and explanation. Further, if a student only gave a low-level response the CC would prompt for more information – asking why? How? The consequence of using guest presenters in this course was that the CC has a greater responsibility to provide opportunities for developing intended student outcomes. These opportunities were not maximised in B&S-V1, where around 70% of the course involved passive sessions. As a consequence, CT learning gains were not as great as anticipated. However, through the introduction of CT intervention tasks, and increase in the amount of discussion and questioning CT outcomes were strengthened.

The use of CT interventions had a positive effect on student CT development. This result is not surprising since previous meta-analyses have determined an overall positive effect of interventions on CT development (such as Pascarella and Terenzini, 1991, 2005; Abrami et al., 2008; Niu et al., 2013; Abrami et al., 2015; Huber and Kuncel, 2016). In contrast, Huber and Kuncel have also discounted the value of general CT training in college, because their meta-analysis findings show that CT develops from college without the need for intervention. However, Dwyer and Eigenauer (2017) contend that Huber and Kuncel are wrong in their assessment, suggesting there is value in both CT interventions and explicit teaching. The importance of explicitly discussing CT to increase student understanding about CT has been highlighted in my study. Students from both course versions associated the course with developing CT abilities and also science literacy skills. Responses to pre-test and post-test self-evaluation of skill level by students in B&S-V2, also indicated they judged themselves to be better at these skills. In the chemistry case study, where CT was entirely embedded in the course, students did not perceive that their CT abilities had improved. Yet, the findings from the B&S course, where there was a significant positive shift in student perceptions of their skills, demonstrate that an explicit CT environment can help improve student confidence in their CT development. Beck and Blumer (2016) found that student and educator perceptions of instructional practices vary if educators do not explain the approaches they are using. Beck and Blumer's findings, in addition to the trends in my research, suggest that the explicit mentioning of CT can play an important role in increased student awareness and perhaps understanding of these abilities. However, Beck and Blumer (2016) also note that there is not yet a reliable way to determine which perspective is more accurate in predicting student learning gains. There is always the possibility that neither is a useful metric of CT learning gains. In the B&S course, student perception and performance results did not correlate; but neither did the CC's expectations about the course and student CT performance. However, Dwyer and Eigenauer (2017) suggest that "the variance in effect sizes observed in extant research on CT may be a function of the difficulties discussed with respect to CT conceptualisation, instructional design and assessment" (2017, p.94). So perhaps it is just a matter of having metrics which are better aligned to instructional and assessment practices.[22]

5.5.3 Future learning opportunities in this course

The aims of this course were to develop a student's capacity to reflect on the interplay between biological and social perspectives and the issues and evidence that underpin them. Based on student perspectives of the generic science skills and CT abilities, the course is meeting the goals of giving students an appreciation of the role and relevance of scientific evidence to society. However, CT performance results were indicative that further training is needed to help produce graduates who are able to use their scientific literacy skills beyond a classroom activity (in line with Flinders University graduate qualities 2 and 7- see Appendix I: Table I.2).

This chapter has already modelled how the multi-tool assessment approach can be used to improve CT outcomes. Modifications made between the first (B&S-V1) and the second version (B&S-V2) of the course led to an improvement in CT performance in B&S-V2. Specifically, there was a reversal of two statistically significant declines in performance, as well as a significant increase in pre and post-test performance for another question (Q2). But despite B&S-V2 having

[22] A discussion of the CAT assessment as a metric for assessing CT in my studies is covered in the next chapter.

increased CT development on some CAT questions, these performance changes only represented an improvement on 20% of the test question, meaning there are opportunities for the course to further improve student CT abilities. Additionally, even in B&S-V2 there were 8 questions where students scored less than 50% of the available CAT points. For five of these (Q3, Q4, Q7, Q11 and Q15), there was a decline in performance between the pre-test and post-test in both versions of the course (however, the magnitude of the decline was smaller in B&S-V2). For the remaining questions (Q9, Q10, Q13) there was an increase between the B&S-V2 pre-test and post-test results despite the fact that less than 50% of available points were scored.

Interestingly, most of these questions with low performance had skills that were targeted by an intervention – suggesting poor transfer between skills developed in a class activity to a close, yet different context. Given this course met most of the conditions recommended for near transfer,[23] it is likely that the previously discussed motivational factors are an important contributing factor to the performance decline. But considering the 2:1 ratio of passive and active teaching sessions in this course, further changes could be made to increase the amount of CT development opportunities. For example, additional CAT-like questions could be added to the course, placing more emphasis on real-world scenarios that require students to interpret and draw conclusions in scenarios where the judgement is context-dependent. However, there are already three CAT-like activities included in the course and the CC likes to vary up the learning activities. Therefore, another option could be to evaluate student work samples using the Adaptive Critical Thinking (ACT) Framework (see Chapter 2), to identify which aspects of CT process students are struggling with. For example: are students able to identify the key problem in the topic they have chosen (ACT element - purposeful querying)? What criteria have they used to make their judgement (ACT element - applying criteria)? Have they used valid evidence to support their claims (ACT element – information processing skills and applying criteria)? Do they have enough background understanding of the topic (ACT element – existing knowledge)? Have they produced a well-constructed argument that addresses the key issue they identified in the purposeful querying stage of the CT process (ACT element – critical reflection)? These types of questions could either be used by the educator to formatively evaluate student work for the purpose of identifying things to cover in class. Alternatively, these types of questions, or the framework itself,

[23] The conditions are: multiple opportunities to practice, instructor modelling of CT and scaffolding of CT tasks, etc. (Merril, 2013; Tiruneh et al., 2014; Tiruneh et al., 2018).

could be used by students when planning and drafting their work. This would enable them to personally reflect on their CT abilities as they solve a problem or form a judgement, and has the potential to increase metacognitive awareness of how they think and how they make decisions and judgements. This use of the ACT Framework would have two benefits. First, it would help students to organise their thinking, and remind them of the thinking steps needed to undertake and produce a well-reasoned perspective in their exposition. Second, the literature suggests there is a positive feedback loop between self-monitoring and CT skills (Ghanizadeh, 2017). Meaning the other benefit of this self-monitoring using the ACT Framework is it will help strengthen a student's CT abilities and vice versa. The ACT Framework could also be incorporated into group discussion activities, as reflective dialogue strategies have also been found to help foster CT development (Erdogan, 2019; Kuhn, 2019).

5.6 Conclusions

This chapter offers an important first insight into an Australia science and society course. It characterised the classroom environment and learning outcomes, and revealed that courses which discuss CT explicitly and critically analyse themes and arguments around science in society can have a positive effect on student perceptions about their CT abilities. However, the results also support previous findings that explicitly mentioning CT is not the most effective approach to improve CT development. They show that the addition of embedded CT tasks alongside explicit descriptions of CT in a course produce stronger CT development outcomes.

This study has continued to address the second broad research question for this dissertation - *What teaching approaches are currently employed by science educators in order to develop critical thinking abilities, and which are most effective?* Through reflecting on two semesters of class learning opportunities, this second case study has shown how the types of teaching approaches used affect the CT development outcomes from an Australian science and society course. As with the chemistry findings (Chapter 4), having a more ALE led to greater CT development outcomes. However, the interpretation of results for this chapter required consideration beyond the standard p-value gatekeeper, and this resulted in a reflection about the role of the expert (and previous findings) in setting the precedents/parameters for evaluating a meaningful effect.

This chapter has further modelled and expanded on a method to explore CT development using the multi-pronged approach presented in Chapter 4. In particular, demonstrating how the findings from this approach, in conjunction with the ACT Framework, can be used to improve student CT development. Importantly, this study has also shown that courses and programs do not need to be completely restructured to increase CT development. A few strategic modifications to the existing classroom environment can lead to stronger CT development outcomes.

Overall, modifications made to the B&S course increased this courses effect on student CT development. However, the magnitude of the performance improvements was smaller than those found for the chemistry case-study. Yet, this B&S course was able to produce perception and disposition changes that the chemistry course did not. These findings increase understanding about the experiences of CT development in Australian tertiary science classrooms, and further contribute evidence about the importance of an active mixed (general plus infusion or immersion) learning environment for stronger CT development outcomes. It has set the scene for the final exploration of CT development in Australian tertiary science classrooms (Chapter 6).

Chapter 6

Comparing Critical Thinking Development in Science

Education is learning what you didn't even know you didn't know.

Boorstin, 1970

6.1 Introduction and chapter outline

This chapter is the third and final in the research series which explores the development of CT in tertiary science classrooms in Australia. Numerous meta-analyses have revealed a range of approaches are used for developing and assessing CT (Allen et al., 1999; Ortiz, 2007; Niu et al., 2013; Abrami et al., 2015; Huber and Kuncel, 2016; Lorencová, Jarošová, Avgitidou and Dimitriadou, 2019), though little literature captures the experiences and outcomes in Australian science classrooms. My educator perception survey findings suggest embedded techniques are most commonly used by Australian science educators (see Chapter 3), which is consistent with Australian policy documents. While the sole use of embedded approaches is not consistent with current best practice recommendations, lack of available data from Australian tertiary science classrooms means it is unclear how this impacts Australian graduates relative to their global peers.

Much of the existing research on CT development is incomplete, despite Facione's recommendations about CT assessment (1990b). Studies are often skills-focussed, and testing regimes remain dominated by multiple-choice and short answer questions (Benjamin et al., 2016). Rear (2019) also questions the validity and the reliability of the sub-scales many of these standardised tests employ. It was important to remedy these issues for two reasons. The first is that many existing CT assessment approaches do not supply the contextual understanding that educators using them want to know. The second reason is to increase educator buy-in. Findings from my educator perception survey (Chapter 3) indicated that many educators do not prioritise formal CT assessment. Therefore, to get buy-in from Australian educators, assessment measures need to demonstrate user benefits.

The collective objective of the case studies was to address aim 3: *to explore how Australian tertiary science educators at Flinders University are developing CT and determine the effect of their approaches using a case-study method that could also serve as a general model for CT assessment.* The previous case-study chapters (Chapter 4 and 5) explored CT development in two different contexts – a chemistry course (embedded CT environment) and in a science and society course (explicit/infusion CT environment). This third chapter addresses two main themes relating to CT assessment. It extends the multi-tool evaluation approach to clarify the trends in the learning gains from the two science courses. It also explores the strengths and weakness of my case-study methodology, including a critique of the CAT instrument.

6.2 Case study comparison methods

The case study process for each course involved observation, perception and performance measures (as described in Chapter 4: Section 4.4). This chapter uses the CAT results (performance measure) and the COPUS data (classroom observation data) from the two previous chapters to explore the trends in classroom environments against the differences in performance outcomes. The information from this cross-study comparison helps address the latter half of Research Question 2 - *What teaching approaches are currently employed by science educators in order to develop critical thinking abilities, and which are most effective?*

6.2.1 Statistical methods

To reveal if there was a particular set of student experiences that resulted in stronger CT development outcomes, this chapter used a between study analysis approach across the three case study data sets. Course differences were explored using descriptive statistics, non-parametric hypothesis testing and multivariate analysis approaches. A summary of these approaches is found in Figure 6.1. To gain an initial sense of variation between the courses, descriptive analyses were used to compare the difference in CAT performance and differences in the COPUS summaries (proportions) across biology and society version 1 (B&S-V1), biology and society version 2 (B&S-V2) and chemistry. Next, the change in students CT performance across the semester were calculated for each course. A Wilcoxon signed-rank test was used to explore for significant differences in the change in performance between the embedded course (chemistry) and the explicit/infusion course (biology and society). There were significant performance differences, between B&S-V1 and B&S-V2 (Chapter 5: Section 5.4.7) and between B&S-V1 and chemistry (see below, Section 6.3), but not between chemistry and B&S-V2 (despite the fact that there was a significant increase in CT performance for the chemistry cohort, but not for the B&S-V2 cohort).

To provide insights into the differences between the courses, multivariate analysis of the classroom observation data was undertaken to seek for aspects of the classroom environment that could be contributing to student CT performance (Figure 6.1). COPUS data from all three courses was transformed into proportions per activity per teaching session, and a sample similarity matrix was generated using a Bray-Curtis Algorithm (Bray and Curtis, 1957). Then, Principal Coordinates Analysis (PCoA) ordinations with unconstrained Pearson vector overlays were generated to determine which of the observation codes correlated (moderately to highly) to the data cloud. These vectors drew attention to COPUS codes that characterise each cluster and helped to distinguish between the similarities and differences between each classroom. These PCoA ordinations and vectors helped display patterns among classroom profiles that can be used to characterise recipes for successful CT development. The spread of the COPUS code vector overlays on the PCoA led to the conclusion that the shape of the cloud seemed to relate to the difference in time spent in active learning and passive learning modes (see below, Section 6.3). The counts associated with active learning codes were then summed together for each course to facilitate comparison between the proportions of active learning time against the change in CAT performance using regression analysis (Figure 6.1). Multivariate analyses were performed using PRIMER (v.6) with the PERMANOVA add-on PRIMER-E, Plymouth Marine Laboratory, UK, (Clarke and Warwick, 2001).

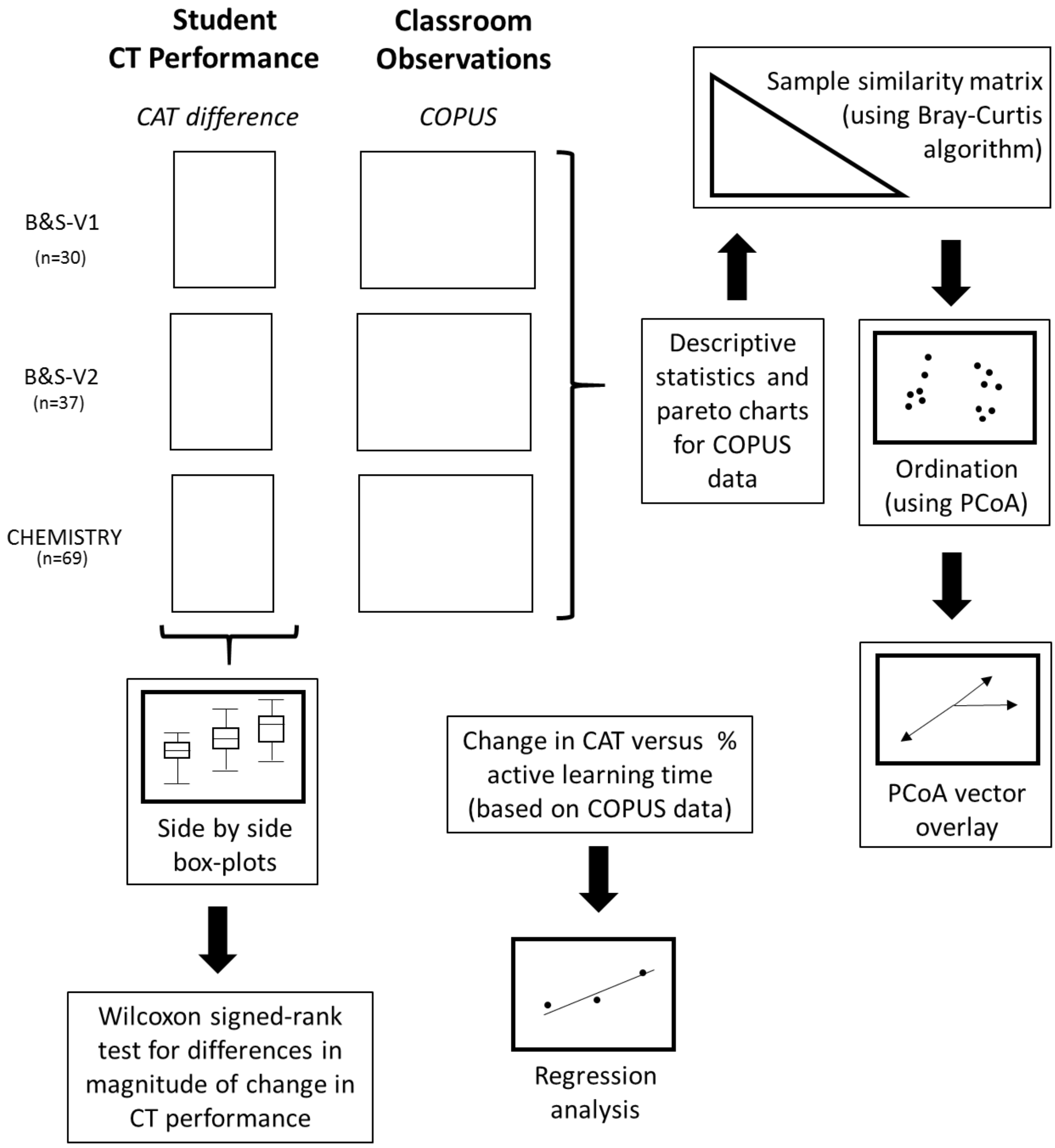


Figure 6.1 Summary of the statistical methods used to make comparisons between the CT performance and classroom observations in the three case study courses.

6.3 Results from the case study comparisons

Comparison of the pre-test post-test CAT score differences for the three courses revealed the greatest change in overall CT performance was in the chemistry course, and the smallest change was for the B&S-V1 course (Figure 6.2A). This increase in the chemistry cohort's performance was significantly higher than the B&S-V1 cohort results, but not the B&S-V2 cohort results (Figure 6.2B). As described in Chapter 6, the B&S-V2 cohort performance change was also significantly higher than the B&S-V1 cohort.

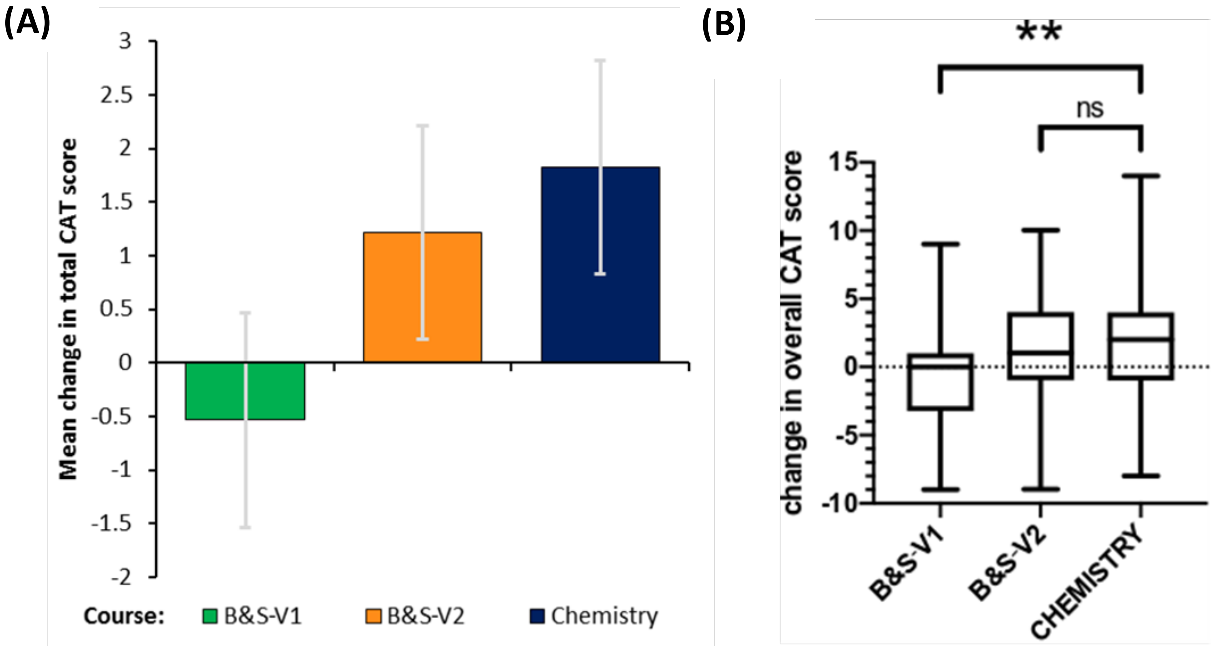


Figure 6.2 A-B Summary of CAT pre-test post-test score differences for the three case study courses. Where (A) displays the mean score change per course (error bars represent ± 1SD); and (B) displays side-by-side box plots of the change in CAT performance for each course.

As Figure 6.3 shows, each course had at least one CAT question where their student cohort improved the most. Students in the B&S-V2 course had the strongest performance gains on 7 questions, however, their overall change in performance was less than the chemistry students. Students from the chemistry course, improved the most on 6 questions (Q3, Q4, Q5, Q10, Q11, Q15 - see Appendix F for a description of the skills associated with each CAT question), and the magnitude of this performance difference per question was generally larger than the other courses (particularly for Q5 and Q11). The B&S-V1 students improved the most at Question 8 (Figure 6.3). All cohort's declined in performance on CAT Q7.

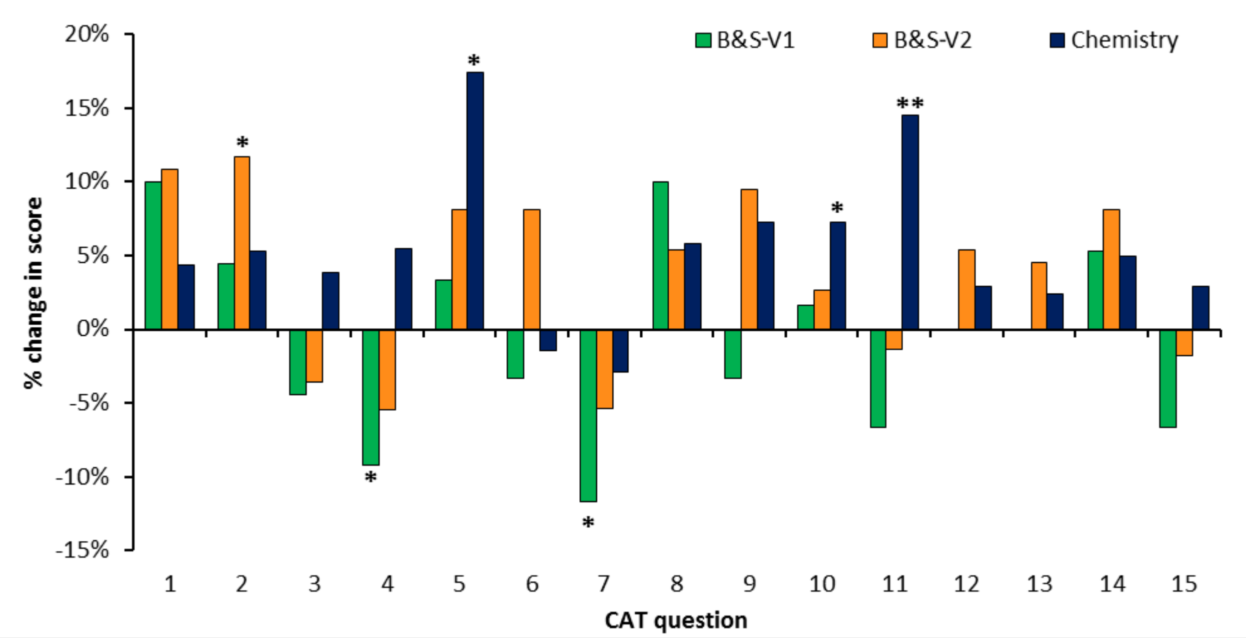


Figure 6.3 Comparison of performances differences (percentage change) on each CAT question for the three case study courses. An asterisk indicates a result where a significant difference lies between the pre-test and post-test performance score within that course and the number of asterisks denotes the level of significance, * <0.05, ** <0.01, * <0.001.**

A number of differences between the classroom environments were evident in the COPUS results. For example, while listening was the most frequently observed student behaviour in all three courses, the total proportion of class time spent listening varied with each course (Figure 6.4). Chemistry had the least (36%), and B&S-V1 had the most (67%). Chemistry students spent the most time working in groups, whereas B&S students spent more time involved in class discussions than the chemistry students did (Figure 6.4). Chemistry instructors spent more time guiding individual student learning (Figure 6.5) than the instructors in either version of the B&S course. However, B&S instructors (particularly in B&S-V2) asked students the most questions during class time (Figure 6.5).

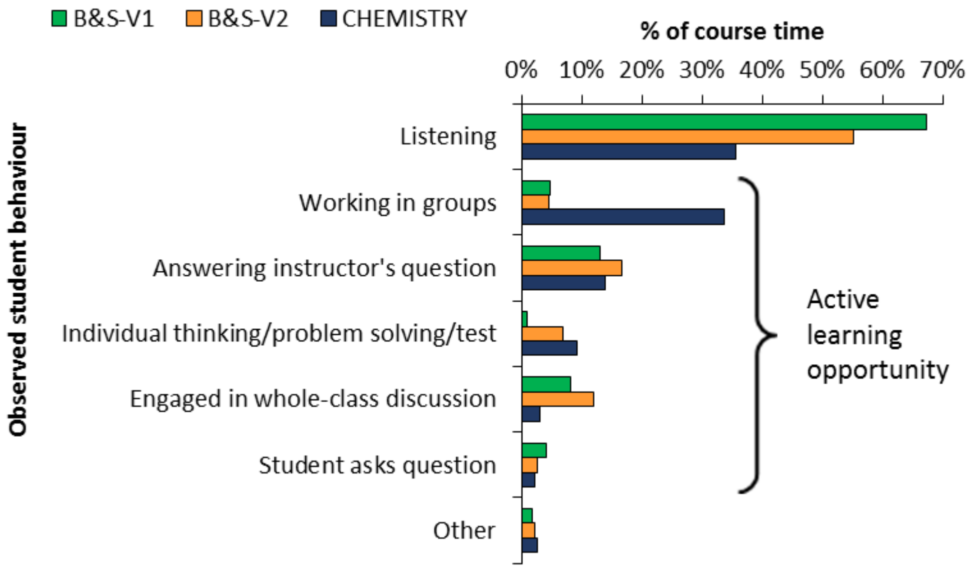


Figure 6.4 Summary of how student time was spent in each of the case study courses.

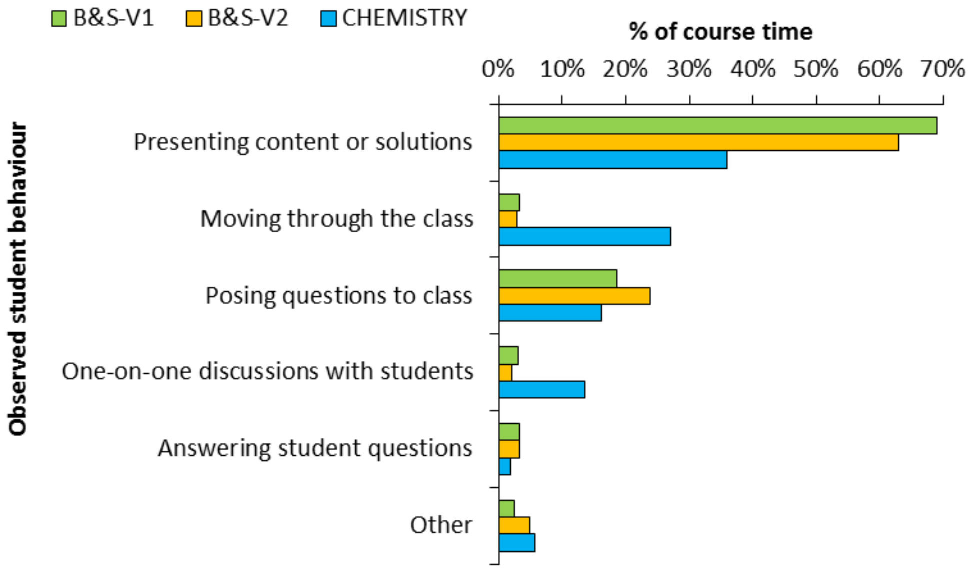


Figure 6.5 Summary of how instructor time was spent in each of the case study courses.

Multivariate analysis of the COPUS data revealed trends by course, by instructor and by session type. 86.7% of the variation in the data was captured by the first two dimensions of the PCoA (Figure 6.6 PCoA1 and PCoA2).

Colouring the PCoA ordination by course-type highlighted some similar parts of all three courses, but there were also some differences (Figure 6.6A). In terms of the differences between the courses, the B&S-V2 data points (orange) form a tighter group than the B&S-V1 data points (green) and chemistry data points (Figure 6.6A). The chemistry data points (dark blue) are spread across the ordination, but form three distinct groups (Figure 6.6A). Figure 6.7 shows these groups represent the three different session types (Figure 6.6B: lecture, workshop and practical). In terms of course similarities, the cluster on the right hand side of the ordination shows a mix of orange, dark blue and green data points) indicating that there were some similarities in the learning experiences offered across all three courses (Figure 6.6A and Figure 6.7). This mixed cluster of courses corresponded to the lecture sessions. The separation of this mixed cluster from other points in the ordination is a consequence of the higher proportion of passive COPUS codes (such as students listening and instructor lecturing) as indicated through the Pearson vector overlays (Figure 6.6B). However, delivery varied within this cluster of lecture sessions. The directions of the Pearson vectors indicate the spread of this cluster correlated to the amount of questions instructors asked. The number of questions asked was highest in the B&S CC sessions, followed by the chemistry instructor sessions, and then the B&S guest lecturer sessions (although there were a few guest lecturers who were more similar to the B&S CC) (Figure 6.7B).

There were also some key differences within each course. For example, the green triangles in Figure 6.6A highlight that while most of the B&S-V1 course was delivered in a fairly consistent manner, there were two sessions that were very different from the rest of the course. Figure 6.7 reveals these particular data points correspond to two discussion sessions. The position of the vector correlations indicates these sessions differ from the rest of the course based on the proportion of time students spent engaging in tasks other than listening. The variability in the presentation styles was smaller in B&S-V2 (Figure 6.6A), but there were instructional differences in how lectures and discussion workshops were delivered (Figure 6.6B and Figure 6.7). Meanwhile, chemistry offered a range of learning experiences and teaching modes (Figure 6.6A and B and Figure 6.7).

There are a few places where the two Pearson vectors correlate highly in approximately the same position on the data cloud (for example, the vectors labelled Lec and Lis in Figure 6.6B). This is a result of including both instructor and student COPUS codes in the sample-similarity matrix, where an action by the instructor prompts a student response - for example, students listening to instructors presenting content. However, the behaviours do not always correspond so both codes have been included in the sample-similarity matrix. For example, students would display the behaviour listening in response to class discussion time as well as during times the instructor lectured.

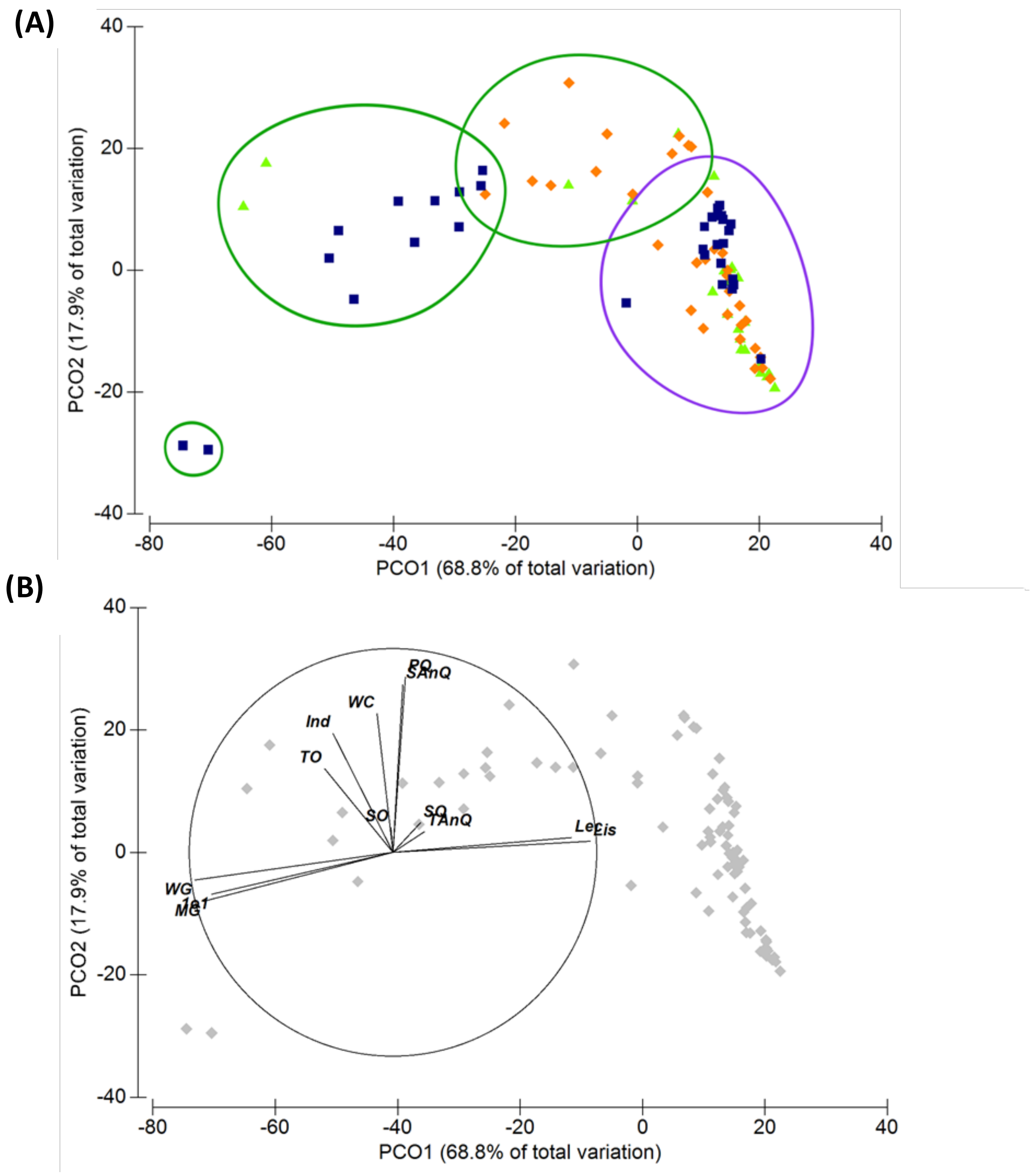


Figure 6.6 A-B PCoA ordination of the three case study courses classroom observation data. Data points in (A) coloured by course type (blue square – chemistry, orange diamond = B&S-V2, green triangle = B&S-V1), where four clusters were identified at a 65% similarity level (circled) where the green circles indicate sessions with a higher proportion of active codes and the purple circle indicating those sessions with a higher proportion of passive codes. Figure (B) is the same PCoA ordination but displays unconstrained Pearson vector overlays, where the vectors show the extent of the correction between the shape of the data cloud and the classroom observation codes.

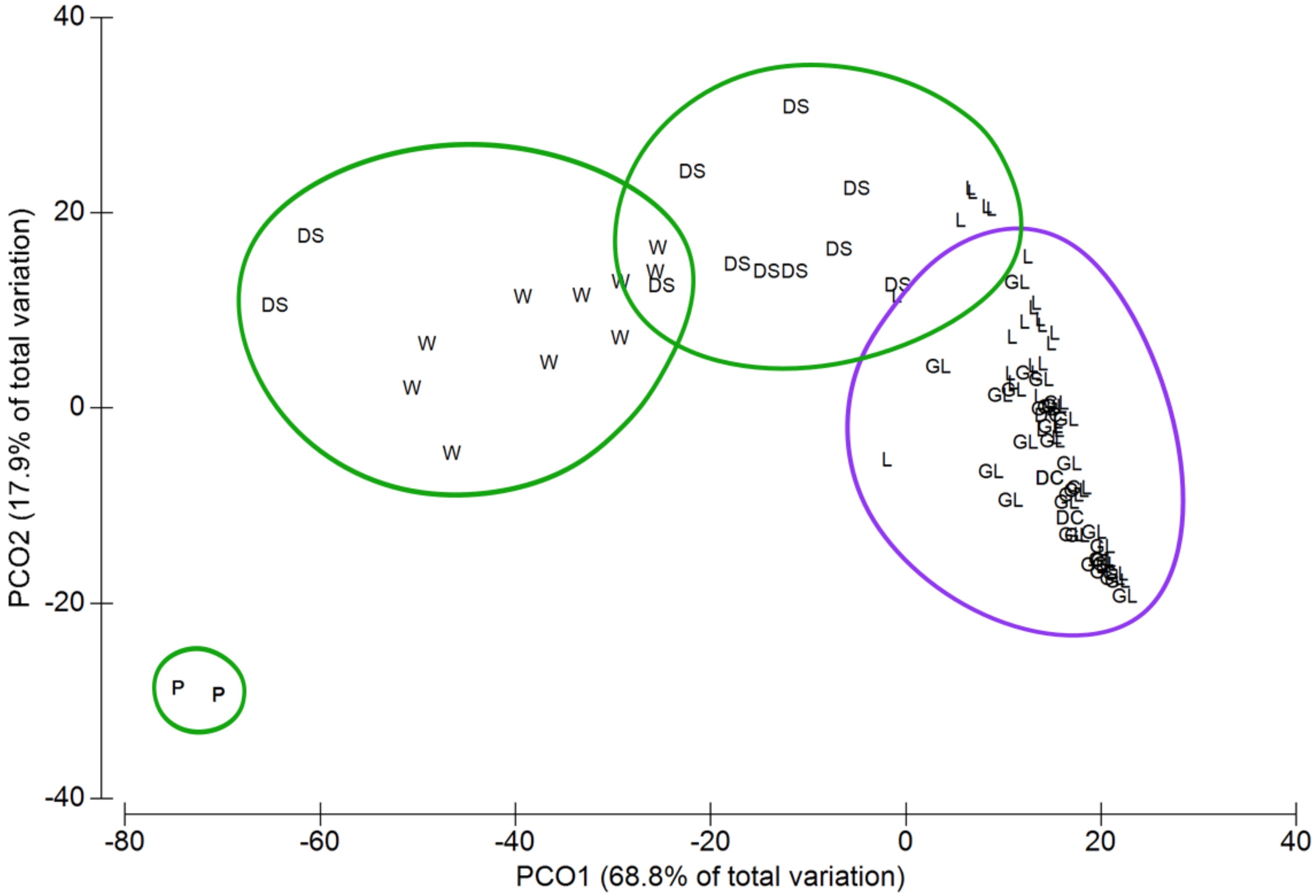


Figure 6.7 PCoA ordination of the three case study courses classroom observation where data points are labelled by session type (DS = discussion; GL = guest lecture; DC = documentary; W = chemistry workshop; L = lecture; P = practical). At a 65% similarity level, four clusters were identified (circled) where the green circles indicate sessions with a higher proportion of active codes and the purple circle indicating those sessions with a higher proportion of passive codes.

The final analysis involved comparing time spent in active learning in each course, against the CAT performance (Figure 6.8). The linear regression model applied to the scatterplot demonstrates a strong positive correlation between the amount of active learning time and change in CAT performance ($r = 0.894$).

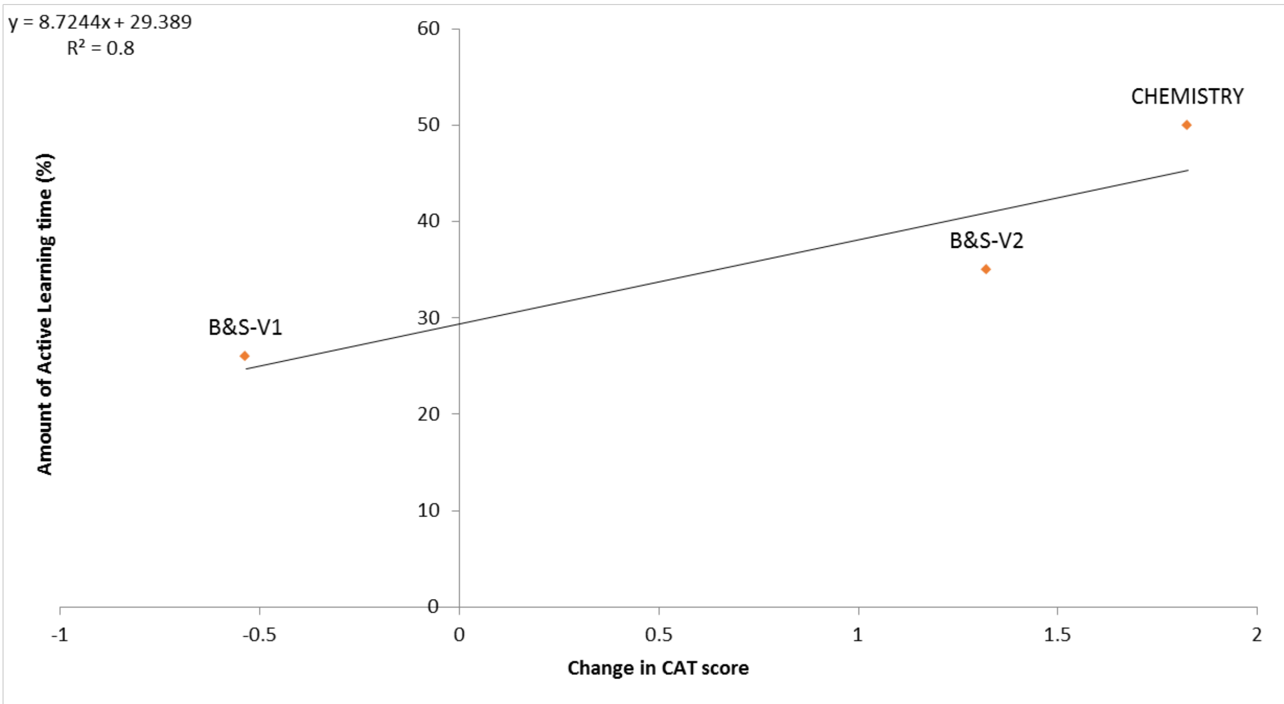


Figure 6.8 Comparison of the change in CAT score versus the amount of active learning time using a linear regression model. Data labelled by course.

6.4 Discussion

This series of case studies were constructed with two goals in mind. The first goal was to add to the understanding about CT development in Australian tertiary science classrooms. The second goal was to construct and explore the value of a novel multi-tool descriptive and performance assessment method to see if it provides a better understanding of CT development. Given the findings from the educator perception study (Chapter 3), which highlighted the embedding of CT within discipline-specific courses, an additional purpose of these case studies was to investigate CT development using embedded methods. This discussion is divided into two sections, the first explaining the findings from the cross-study comparisons, and the second discussing the general case study methodology.

6.4.1 Discussion of the effects of the two science courses

This particular study compared the classroom experiences and CT learning gains from the two previous case study chapters, to better understand CT development outcomes that occur in Australian tertiary science classrooms. I explore the differences in CT development outcomes between a general scientific literacy course (explicit/infusion CT environment) and a discipline-specific course (embedded CT environment). Multivariate analysis methods were used to seek for patterns across courses. This was done to determine whether particular sets of student experiences produce a stronger effect on CT development (in terms of overall development and/or a particular skill development).

Based on findings from previous studies (Abrami et al., 2008, 2015) the *a priori* expectations were that courses which discuss CT explicitly would have a greater effect than courses that embed CT among curriculum content. Unexpectedly, it was found that an embedded learning approach had a statistically stronger effect on CT development than a course with more explicit/infusion CT teaching. However, there are three reasons I caution against the conclusion that discipline-specific pure science courses (i.e. chemistry) are more effective at developing student CT than a more general scientific literacy-based course (i.e. B&S).

Firstly, the embedded course included proportionally more active learning opportunities (such as group work and time for individual thinking) than the explicit/infusion course. Thus, the increase in opportunity to actively learn was likely the driver of improved CT rather than the embedded nature of the course. This is supported by evidence from B&S where the interventions in the second year increased the proportion of active learning time by around 10% and the trend was improved CT performance over and above the B&S-V1 version. This trend of active learning correlating to increasing CT performance is comparable to the findings of Styers et al. (2018).[24] There are also a number of other articles which suggest that ALE lead to stronger learning outcomes (Bonwell and Eison, 1991; Haak et al., 2011; Kim et al., 2013; Lumpkin et al., 2015; Eddy and Hogan, 2014; Wright et al., 2016; Bezanilla et al., 2019).

The second reason caution should be taken about drawing the conclusion that the discipline-specific science courses are more effective than general scientific literacy-based courses

[24] This study did not measure the proportions of active learning time in their courses; instead academics self-reported their estimates of the proportion of their course that was active.

is that the embedded course (chemistry) had many more contact hours than in B&S (5 hours compared to 3 hours a week). This facilitated more regular and increased engagement in behaviours that help develop CT in the chemistry cohort than in the B&S cohorts. The literature indicates the amount and regularity of practice is important for CT development and transfer (van Gelder, 2010; Heijltjes et al., 2014; Wall, 2015). This pattern of regular practice leading to stronger outcomes is also supported by the findings in B&S-V2 where there was an increase in the frequency of CT practice, and a stronger CT outcome. Therefore, it was the quantity and frequency of opportunities supplied by each course environment, rather than an attribute strictly offered through a pure-science classroom, that likely amounted to differences in the CT outcome.

The third reason to be cautious about concluding that pure science environments lead to stronger CT outcomes than general scientific literacy courses relates to the quantity of CT related learning gains affected by each course. While chemistry had a bigger overall change in performance pre-test to post-test, B&S-V2 had a positive effect on a greater variety of CAT questions, and therefore CT skills. In addition, students from the B&S course seemed more aware of what CT is (including changes to their CT abilities). So even though the magnitude of the effect was smaller in the B&S course, the impact on student CT skills, perceptions and dispositions was greater. This trend fits with previous literature findings that a mixed teaching approach (general plus infusion or immersion) leads to stronger CT outcomes (Abrami et al., 2008, 2015). This is also supported by the results from the comparison between the overall performance differences resulting from each course. This analysis revealed no statistically significant difference in the extent of the change in performance between B&S-V2 and chemistry, even though there was a significant difference between B&S-V1 and chemistry.

These results demonstrate a fine balance between the inclusion of theory and practice. In a recent review, Lorencová et al. (2019) noted that too little practice can result in no or low effects. Additionally, when minimal or no CT theory is provided, students do not develop a grasp of CT methods, which can inhibit successful learning gains. Heijltjes et al. (2014) also found that both explicit instruction and opportunities to practice CT-related skills were necessary to have a positive CT outcome. The results from the two versions of the B&S course also demonstrate this. In B&S-V1, there was some CT theory but not a lot of opportunities to practice CT – and there was no effect. However, in B&S-V2 when the amount of practice was increased and scaffolded among theory, CT perceptions and performance increased. Therefore, the takeaway message from this set of studies is that more frequent engagement in problem-solving tasks, in an active learning environment (ALE), facilitates stronger CT performance outcomes. This is consistent with findings from other studies on CT (Holt, Young, Keetch, Larsen, and Mollner, 2015; Sgambi, Kubiak, Basso and Garavaglia, 2019) and also on learning gains from active class environments (Jensen, Kummer, and Godoy, 2015). In particular, Holt et al. (2015) found that ALE are the best model for explaining gains in CT performance development. However, my case studies have also shown is important to include some explicit training about CT if an educator is interested in improving student understanding of CT methodology, as well as student's awareness of and dispositions towards applying their CT skills.

6.4.2 Lessons from classroom-based research

To improve outcomes such as CT development in science classrooms, time needs to be devoted to researching it, reflecting on it and researching it again (Shipley et al., 2017). However, education-based research is subject to a number of pitfalls, especially when you compare it to science where many controls can be implemented, and experiments rerun. If using a true experimental design, educators face the logistical problems of trying to manage a control group and one or more treatment groups within a course or classroom. There is also the ethical dilemma of withholding the improvement from the control group, particularly when they anticipate a positive learning outcome from the new pedagogy. Sample size (Kranzfelder, Lo, Melloy, Walker and Warfa, 2019) and engagement and student motivation are key issues (Bensley et al., 2016; Finney, Myers and Mathers, 2017; Simper et al., 2019). Generalisability of results and willingness for STEM educators to take on new approaches is limited due to the context-specific nature of teaching and learning (Lund and Stains, 2015; Shadle, Marker and Earl, 2017; Bathgate, Aragón, Cavanagh, Frederick and Graham, 2019a). Yet Gouvea (2017) notes even small-scale studies can provide important insights into the underlying mechanisms of interest. These are all general issues in education research; add in the complication of trying to assess a construct that has multiple definitions and the range of possible issues with trying to determine an effect grows.

The rationale concerning the tool choices for my case study methods were explained in Chapter 4 (Section 4.4.1). The approaches were chosen to help supply a solution to the sheer number of options between definitions, development approaches and assessment tools, which combined with inadequate professional training and support, meant that educators would benefit from an adaptable method to investigate CT development. I wanted a series of instruments that could bring understanding to the different facets of CT as well as the classroom environment. A balance was struck between using standardised measures and existing survey instruments and adapting them for my context. In terms of a holistic stance, my method was a little weak on the dispositional aspects. The goal was to keep it simple for educators with no formal CT training and options such as Halpern's disposition assessment are time intensive as well as quite subjective to mark. The CAT is also time intensive to mark; but is less subjective and the value (training, discussion and feedback) educators get from this instrument made it worthwhile. However, the CAT process has a few issues users should be aware of before deciding if it is the right tool for their classroom. These issues, as well as a discussion of my observation tool and my survey instrument, are covered next.

Assessing CT using the CAT

Critical thinking is a perplexing construct to assess. There are a number of instruments available to use; however, it is challenging to find one that is fit for purpose. Question choice is important, and where possible MCQ instruments should not be used to assess CT (Bassett, 2016; Fukuzawa and deBraga, 2019). The CAT best met the requirements I set for my research. These requirements were based on recommendations from the literature (open-ended questions, science-like question themes, and transparent scoring process). To date, there are 47 publications that have cited one of the very first CAT publications (Stein et al., 2007). However, my case studies on CT development are the first to use the CAT in Australia.

The CAT is somewhat similar to the CLA, which has been accused of being too Americanised (Aloisi and Callaghan, 2018). Given this, it was worth comparing the Australian students against their American counterparts (freshman) to determine potential performance differences. This comparison helped with checking if Australian students understood the test phrasing since it used imperial rather than SI units. Australian students outperformed their freshmen equivalent in both the pre-test and post-test (Chapter 4: Section 4.6.2 and Chapter 5: Section 5.4.7), suggesting there were no obvious issues with them understanding the test. However, there did appear to be some cohort differences. The Australian freshman students started off with higher performance results that the US freshman students, likely because the South Australian education system includes research projects[25] and CT embedded as a general

[25] The research project is unique to the South Australian senior secondary student experience. It would be interesting to compare the CAT performance results across Australia to see if it makes a measurable difference in students CT abilities as a future direction for this research.

capability through primary and secondary schooling. The magnitude of the performance differences was smaller in my data sets than those published. This may be a result of the pre-test results being higher; meaning the influence of first year of university on CT development could be smaller than what American freshman experience. However, the role of motivation on the magnitude of change has also been found to be an issue with test like the CAT (Bensley et al., 2016; Simper et al., 2019).

Previous studies have suggested that there is a positive small to moderate correlation between student motivation and test performance on standardised CT tests (Simper et al., 2019). This is important to consider since the Australian students were not incentivised to complete the CAT whereas in many of the US studies American students were incentivised. Students completed the post-test during week 12, a time when most courses have final summative assessments due so there were competing priorities for student attention and disruptions to students' regular sleep patterns. The CAT questions being the same in the pre-test and post-test means students could have been less interested or motivated to provide a thoughtful response at the post-test (especially since there was no incentive for completing the test, and the post-test was conducted at the end of the semester when student priorities are on summative assessment tasks and exams). However, the use of the same pre-test and post-test was unavoidable because at the time of my data collection, only one version of the test was available. Therefore, lack of incentivisation, test-timing and the same pre-test and post-test are all factors that could help account for the smaller magnitude of change observed in my courses compared to those seen in US classrooms. But upon receiving the CAT some unforeseen issues were identified, and during scoring training and scoring the tests further issues became evident.

The first hurdle related to terminology differences. This affected the delivery and assessment of student performance in a number of ways. Upon receiving the first batch of tests it became evident that the use of US metrics rather than standardised ones could be a stumbling block for Australian students. It was not possible to change the wording of the test, so instead a glossary of terms indicating the metric equivalent was provided to help overcome this. Students were given a glossary to help manage the terminology differences. However, the language difference and need to refer to a glossary adds another cognitive layer to student interaction with the test before thinking and constructing their answer. Given the higher performance of the Australian students, it does not seem like this added cognitive load was detrimental to their performance. However, it is worth considering how the different education systems, cultural differences and rules relating to the natural environment (since there are questions which draw on understanding about national parks) would change how an American student and an Australian student would engage with the test content.

The second issue with the terminology was the phrasing of the questions. Often the language used in the question was very general and this seemed to present a stumbling block for the Australian science students. For example, it became evident when scoring the first batch of Australian student tests that some had interpreted the word 'describe' in line with a science perspective (providing a summary) when the scoring guide was actually looking for a response that included 'description' and 'interpretation' (with emphasis on the explanation why). The questions on their own were fine; however, when combined with the scoring guide it appeared there were hidden or nuanced expectations (see Appendix F for a sample question and scoring explanation). Perhaps this was not identified as an issue because many US degrees include studying English, Math and science before specialising, so discipline-specific interpretations of terminology may not be as prevalent. However, I found that the phrasing was open to subject-specific misinterpretation by students in my cohorts.

Aside from these two issues, which would mainly affect my cohorts, there were two further issues identified with the scoring guide which would affect **all** CAT results. Comparing the question phrasing against the assessment expectations outlined in the guide revealed some issues with alignment between the two.[26] In particular the scoring guide sometimes assesses beyond what the question would ask (refer to Appendix F for an example using a disclosable sample question). While scoring in this manner provides an opportunity to reward students who think beyond the question (deeper than others), it disadvantages those who follow instructions. For example, there was a question that asked students what additional information was needed to evaluate a conclusion. The mark scheme for this question only awards points if the student explained how each alternative they generated contributes to evaluating a particular conclusion (with no points awarded for generating relevant suggestions about the types information needed). This issue could be rectified by awarding points for the generation of relevant alternatives, with additional points awarded for explaining how they would be used to evaluate it. Some of the scoring guide did this, but other parts did not – only awarding a point if the explanation was

[26] *Note:* Individually, both the questions and the scoring guide seem appropriate for measuring CT.

provided, so the scoring guide also seemed generally inconsistent in its approach. The conversations I had with the science educators involved in my research indicated they also shared the opinion that the main issue with the CAT process was the scoring guide rather than the questions. We also reflected on the nature of language and how scientists interpret words differently to non-scientists and felt this could have contributed to our perception of the misalignment.

A second issue I had with the scoring guide was the lack of consistency in awarding points. For example, some questions awarded points when students did not follow the question instructions correctly (i.e. if the question asked students to select one option and the response included two options there were circumstances where the students could still receive points). What made it seem particularly inequitable was that the scoring criteria for the follow up question were constructed such that students who had previously picked more than one option could actually be awarded more points than students who had followed the instruction and picked one option. A further issue was point deduction approaches were inconsistent. For example, one question deducted points for selecting too many options, another question did not. One question included point deductions for wrong choices, another only included point deductions for wrong choices if they had selected too many options (i.e. students who followed instructions about the number of choices required but made some incorrect choices did not have points deducted for their wrong choices). Some of the issues with the scoring and question alignment were raised when the primary researcher attended the scoring training in 2014, and while the CAT director entertained the discussion ultimately the comments were disregarded (likely because the institute does not want to adjust or invalidate existing data). However, it seems invalid, or at least not in the spirit of critical thinking, to continue using a method with identifiable yet rectifiable weaknesses.

Changes could be made either to the test or the scoring process – depending on which approach would make it easier for the CAT creators to distinguish between the previous version and updated version. One suggestion is that modifications could be made to the scoring guide with some retrospective analysis applied to the previous scores to make the results comparable. For example, if the scoring guide was updated to increase alignment, the creators could work out how much it changes the average CAT score by (or how much it changed particular question scores or CT skill sub-scores) then this could be supplied as a guide for interpreting previous CAT results without entirely invalidating the data. Updates to the scoring guide (noting the average change to score) would also allow them to continue to make relevant comparisons to their existing data pool, whereas modifications to the questions would impact the comparability of old and new data. However, this task was beyond the scope of my current research.

Ultimately, I still believe this tool choice was an appropriate one - as the issues with scoring alignment could exist with other assessments and would be harder to detect (since other instruments keep the scoring process hidden). The CAT scoring process gave teaching staff deeper insights into student growth than an MCQ score or sub-score. Further the overall benefits of the CAT, particularly the capacity for developing interventions, means it remains a useful tool for assessing CT in this research and beyond.

Student perception survey

Each of the case studies revealed differences in student judgment of their skills and performance. In B&S, perceptions changed markedly but performance only a little, whereas in chemistry performance changes were greater than perception changes. This is not surprising given students are often inaccurate in their self-assessment of their abilities (Armitage and Connor, 2001; Brown, Andrade, and Chen, 2015), especially CT (Davies and Barnett, 2015; Rayner and Papakonstantinou, 2018). However, the SALG instrument has been found to have more success than most (Seymour, Wiese, Hunter, and Daffinrud, 2000), and it was the SALG items that the students were more accurate at reporting on. Beck and Blumer suggest that perceptions of instructional practice between educators and students can differ if instructors have not been explicit about their approaches (2016, p.15). This trend was evident in the chemistry course used embedded approaches where students were not as good at recognising the CT-specific learning opportunities as they were at identifying the more science-specific learning outcomes. However, there are some performance-based implications from this. It suggests that while the routine of thinking critically is familiar for chemistry students (and CT ability-wise they have improved), their conscious recognition of their capacity to perform these mental processes has not experienced the same benefit. Which implies their capacity for transferring these skills outside of the specific chemistry context could be impeded.

The challenges of making classroom observations

The observation methods for this study included using the COPUS, EQUIP and QSRLS protocols. The COPUS and EQUIP protocol are structured protocols that have been useful for informing numerous science educators about their classroom environments (Cian, Marshall and

Qian, 2018; Stain et al., 2018; Kranzfelder et al., 2019). But conducting the observations was a time-intensive process, which is problematic given that educators are already time-poor (Carbone et al., 2019). Studies by the COPUS creators have reported using between 1- 3 observations to obtain insight into an educator's instructional approaches (Smith, Vinson, Smith, Lewin, and Stetzer, 2014), whereas publications by the EQUIP creators do not supply values for how often the EQUIP tool should be applied. But given the exploratory nature of this novel tri-tool assessment method, I chose to observe more classes to be able to generate a full picture of classroom opportunities. Future directions for this research could include working out the minimum number of observations to produce an accurate understanding of the classroom environment for each course. However, until this work is undertaken, to make this assessment process more achievable for time-poor educators, observations could also be undertaken over a number of years. Alternatively, Evenhouse et al. have come up with a video-based approach to enable "continuous and flexible coding" (2018, p.98) of active-learning, blended and collaborative classrooms which offer a method for exploring classroom environments in an educator's own time. This allowed educators to observe a particular aspect of the course and make progressive changes. As the B&S-V2 results demonstrated that small, but targeted, changes to the amount of active learning time and CT related-tasks were able to produce a meaningful change in CT development. Another option is to have peers assess their classroom using the COPUS instrument, or other observation tools such as the Practical Observation Rubric to Assess Active Learning (PORTAAL) which is specifically designed to assess active learning in science (Eddy, Converse, and Wenderoth, 2015). PORTAAL was not available at the time of designing my research protocol. However, since my findings and others (i.e. Holt et al., 2015) demonstrate the importance of ALE for fostering CT development, if I undertake similar studies in the future I would trial both the PORTAAL (Eddy et al., 2015) and *Freeform* (Evenhouse et al., 2018) instruments.

Investing time in observations is worthwhile, irrespective of the observation tool or if an educator chooses to focus on part of a course or the whole course. Many studies report meaningful insights into the classroom (Amrein-Beardsley and Popp, 2012; Smith et al., 2014; Teasdale et al., 2017), and often these lead to instructional change (Viskupic et al., 2019; Dillion et al., 2019). In my case studies, the observations helped uncover inaccuracies between perceptions about expected learning outcomes and learning opportunities and the actual learning opportunities and outcomes. For example, based on previous findings about CT, it was anticipated that the B&S course would have improved CT development and further, it would have a greater effect than chemistry (where CT opportunities are embedded). Yet the reverse was found. This trend boiled down to the more active learning time the greater the CT development. However, determining the fraction of course-time required was outside of the scope of my research and requires data from many more courses over many years. Therefore, additional research should also be carried out to clarify the optimum amount of active learning opportunities required for CT development. Even within my findings, replication studies are needed to further clarify the trends within each course to take out the potential differences in CT development being related to cohort effects rather than the classroom environment. A further limitation is that I focused on one course per student, yet students were concurrently participating in other courses. Therefore, the experiences outside the classroom I measured could also be contributing to the findings.

There were plans to explore CT development by degree program, however, the resulting data sets were too small to explore for variations that may be related student study programs. In addition, classroom observations were limited to chemistry and biology and society – not all the first-year courses study participants were enrolled in. So even if the data sets were sufficiently sized to explore by degree program, additional observation data from all courses study participants were concurrently taking would be needed to capture the sum of a student's experiences across the semester. This would require significant resources and the implications of such a study would be better suited for exploring institution level-change or discipline-area change (Teasdale et al., 2017) rather than individual courses, and thus was beyond the scope of my work. However, this further indicates the need for more research to determine the number of observations required to obtain a full sense of the course environment.

There were plans to extend the observation protocols by making judgements about the cognitive level of questions asked to students in each class. To achieve this, data was collected on the phrasing of the questions asked to students, with the intent of analysing the nature of the thinking demand and the terminology used when asking students questions. It was anticipated that reflecting on these cognitive opportunities (per question) would provide deeper insights into the ‘questioning' COPUS code, which could then further enhance understanding of the performance information collected through CAT. But sometimes the pace of questioning in class made it hard to judge the thinking opportunity each question provided. To address this issue, the wording of the question was recorded (instead of making the judgements live), and then after class, a cognitive level was assigned. However, these were not used in the final analysis as the pace at which questions were asked in some of the live classes (where recorded data was not available) meant that it was challenging to accurately record the phrasing of the question. This represents another opportunity for further exploration, as the findings from Abrami et al. (2015) and experiences from these case studies suggest that language plays an important role in student awareness of CT, the transferability of CT and therefore of CT development.

6.5 Conclusion

The case study chapters have employed a novel methodology that consists of a barrage of tools designed to help capture learning opportunities and learning gains. The measures were able to reveal what students improved in, areas for development (both in term of CT skills and classroom experiences), what students perceive about their skill levels, and what they thought they experienced in the course. Other studies that have explored CT development often focus on perception and/or performance results. However, this study has shown that observing the classroom environment provides important insights for unpacking why students may or may not be achieving expected learning outcomes. The classroom observation using the COPUS tool helped classify the activity types within the course, and provided an in-depth summary of the experiences within each of these course components.

This chapter generated understanding that contributed to *aim 3 - To explore how Australian tertiary science educators at Flinders University are developing CT and determine the effect of their approaches using a case-study method that could also serve as a general model for CT assessment*. This chapter has revealed the power of using transdisciplinary methods to generate understanding about CT development. In particular, this last study demonstrated that the novel multi-instrument approach was able to reveal deeper insights into CT development, than any of the individual assessment measures on their own. The specific CT related outcomes could not have been discerned from the observations of the learning environment and the student perceptions alone. Similarly, specific awareness of the classroom behaviours and knowledge modes, in addition to an understanding about student perceptions of the environment they experienced would not have been captured through the CAT instrument alone. Importantly, the use of this methodology generates usable knowledge about a course, revealing gaps in student understanding and experiences that can be targeted for further CT gains. This understanding, in concert with targeted interventions (such as CAT-like questions or the use of my ACT Framework), can be incorporated into future versions of the course to deliver stronger CT development outcomes. This research has also contributed understanding about the complexity of CT assessment, providing a critique of the CAT scoring approach that has not been published before.

This chapter has also provided the answer to the latter half of my second broad research question: *What teaching approaches are currently employed by Australian science educators in order to develop critical thinking abilities, and which are most effective?* The addition of multivariate statistics helped to further clarify the instructional differences in learning environments of the three courses. This helped to generate new understanding about the instruction styles Australian science educators' use, and which approaches lead to the strongest critical thinking outcomes. Specifically, ALE where CT is either embedded or infused explicitly, lead to the greatest CT learning gains. This research has shown that CT development can be achieved when only a portion of the course, such as a workshop or practical session is active.

PRIMER 4

I started this doctoral journey with the naïve assumption that because critical thinking (CT) was a term touted throughout my education, that ideas about the construct would be well resolved. However, this belief could not have been more wrong. Defining and understanding the development of CT were, in fact, wicked and persistent problems in education.

CT development is a worthy goal that educators should continue to aspire to. However, as discussed in my literature review, the breadth, depth and conflicting ideas in existing CT theory against a backdrop where educators are typically lacking time and training presents multiple barriers to achieve this goal. As an individual, I acknowledge that am unlikely to ever be equipped to help every educator overcome their time and training challenges – however, I was also not satisfied to leave this problem for others to solve. I recognised that I could apply my skills as a scientist to synthesise and increase the accessibility of theory, in addition to formulating a more substantive approach to exploring CT's development.

As demonstrated over the last three chapters, unpacking CT development is complex. If the lack of a consensus definition did not make it complex enough, the variety of perspectives on development and the range of assessments further complicates this area. My case studies step out of these limitations to explore current practice, and there are seven key findings relating to CT development and CT evaluation that resulted from the chemistry and biology and society case study research.

Concerning CT development, the takeaway messages are:

- CT can develop when embedded in a discipline-specific course when the class environment is structured for students to be engaged in active learning.
- Explicit teaching of CT theory does not necessarily lead to improvements in CT skill development, particularly if no or limited opportunities are provided for practice.
- Active engagement with tasks that require students to employ CT improves students CT abilities, irrespective of the learning-context being discipline-specific or general. However, the literature suggests that the magnitude of this change is greater in a mixed (explicit and infusion or immersion) environment (Abrami et al., 2015).

- When you are an instructor that presents all your own content you have more control over the way that content is presented. This can make it easier to ensure that activities which target desired learning gains are incorporated into the classroom regularly.

Concerning CT evaluation:

- A multi-tool approach, while time-consuming, generates a more thorough picture of the classroom environment and learning outcomes. It enables a stronger diagnosis of opportunities for improvement.
- Context is very important. There is no standalone perfect CT assessment, and should never be a one size fits all CT assessment unless it is an adaptable tool. The most important things for an educator to think about when choosing a CT performance measure are – what do I think CT is and does this test assess CT in a way that aligns to my perspective?
- Student motivation also plays a big role in whether the assessment results are an accurate representation of their abilities.

As I reflect on my university journey, I have started to ponder about how I have arrived at a place where the findings of my research could inform education policy and practice. My training as a scientist gave me an appreciation of using knowledge and processes from other areas of science to explore new avenues. It also showed me the value of incorporating and adapting existing methodology to solve complex problems. However, it is through my doctoral journey that I have come to understand that this approach can require the reconfiguration of ideas because most scientific knowledge tends to be rooted in the Physics paradigm of science. A state where thinking, terminology and theory tend to be developed specifically for the immediate context (intrapragmatic knowledge) rather than producing content which can be more widely understood and applied beyond the immediate discipline (metapragmatic knowledge). My view of knowledge needed to change if I was going to be able to contribute understanding that was more widely applicable. I needed to think more about the movement and the sharing of knowledge outside of the frame of science. The role of context became clear through my explorations of CT theory and in the findings from the educator perception survey. I noticed how context shaped the CT stance as well as the associated development and assessment approaches. I recognised that to achieve the transcendency associated with transdisciplinarity (McMicheal, 2000), I needed to think about complexity, context and connection (Montuori, 2019). I needed to become what Steger terms "a reflexive nomad" (Steger, 2019 p767), someone capable of moving around and beyond literature and perspectives from different disciplines to ruminate, unite and grow understanding that would usually be explored in isolation. I needed to think critically about critical thinking.

The various chapters in my thesis have outlined a theoretical framework (the ACT Framework), as well as novel methodology and data to increase understanding about CT. This set of case studies marked the final step in my transdisciplinary journey from thinking like a scientist to thinking like a nomadic science educator. Having already been involved in science teaching and science education research before I started this doctoral research, I thought I already had a healthy appreciation for the nexus between knowledge generation and application, and between student engagement and student learning. But before designing my case study approach I had not put much thought into the delicate balance between of pedagogical knowledge alongside the discipline content knowledge, cognitive skills and learning outcomes. This part of my transdisciplinary journey has seen the merging of knowledge and methods from science, statistics, education and social science in order to help increase understanding about the wicked problem that is CT development in Australia. The power of this collective approach has been demonstrated through all the data chapters in this thesis, however, the cases studies were a culmination of the multi-pronged approach. The combination of assessments, statistical methods and theory demonstrated that this method was not only able to evaluate the effect of a course on CT development, but it was also able to discern likely reasons why the course produced that effect in addition to ways to improve it. Further, through comparisons across three courses, it became even clearer that active learning environments lead to stronger CT development outcomes. Collectively, the set of case studies have shown the value of transdisciplinary approaches to address problems in education.

Conclusion:

From concept to classroom

The knowable world is incomplete if seen from any one point of view, incoherent if seen from all points of view at once, and empty if seen from nowhere in particular.

Shweder, 2003

CT: A renewed focus

Critical thinking (CT) is recognised as an essential 21st-century skill by education systems throughout the world (Greiff, Niepel and Wistenberg, 2015; Oliver and Jorre de St Jorre, 2018; Sellars et al., 2018; Bezanilla et al., 2019; Doeke and Maire, 2019). The recognition of CT's importance has been ongoing (Ennis, 1962; Facione, 1990a; Giancarlo and Facione, 2001; Pietrzak et al., 2018). Jungwirth and Dreyfus (1990) express that critical thinking (CT) development "has been regarded as one of the most essential objectives of science education for more than 100 years" (Jungwirth and Dreyfus, 1990, p.42). The American Philosophical Associations 'consensus' stance on CT (Facione, 1990a) remains one of the most highly cited conceptions of CT and has received renewed attention, with the Delph Panel's report citation rate more than doubling over the last five years. Yet, there are still calls for consensus. Reasons for this were discussed in the literature review which highlighted that terminology is used inconsistently (Cuban, 1984; Edwards, 2007; Sanders and Moulenbelt, 2011), with the same idea described using different language, and different ideas described using the same language. In addition, definitions are not always aligned to the context that they are used (Norris, 1992). Basically, theorists to date have generated an abundance of information, but not much consistency.

Given the state of the conception of CT, it was not surprising to find numerous reports that educators are not always clear about how to include or prioritise CT in their classrooms (for example, Paul et al., 1997a, 1997b; Shell, 2001; Black, 2009; Choy and Cheah, 2009; Aliakbari and Sadeghdaghighi, 2012; Kowalczyk et al., 2012; Stedman and Adams, 2012; Moore, 2013; Alwadai, 2014). Even now, CT is still not part of educator training (Lorencová et al., 2019). This is problematic since an educator's assessment choices are strongly influenced by their past experiences (Leming, 2016). Consequently, the purpose of this research was to provide some clarity and resources to address these ambiguous areas, through evidence-based explorations of classrooms to elicit the perceptions and actions of educators when it comes to incorporating CT. It does this by exploring CT in Australia through three lenses – CT conceptions, CT development and CT assessment.

This doctoral research started with two main aims – to explore Australian educator perspectives about CT; and to explore actions and outcomes from tertiary science classrooms. However, this expanded to a third aim during the process of gathering content for my educator perception survey (Chapter 3). It became clear the variety of ideas in the literature could make it challenging for educators to find a suitable CT definition for their classroom. As an outcome, clarifying the construct of CT for educators became the first aim of my research; however, I worked on this taxonomy concurrently while addressing the other research goals. An Australian focus was particularly important because there are few previous studies from this context. Additionally, the Australian policy-based recommendation of embedded CT development approaches and assessment conflict with ongoing suggestions about pedagogies associated with the strongest CT development outcomes.[27] The fact that Australian policy recommendations are at odds with current understanding about best practice means Australian educators face additional barriers to achieving CT development beyond issues with support and training. Therefore, insights into Australian classrooms are essential to generate understanding about how to best support these educators.

The underlying drive behind all aspects of this doctoral research was to generate mobile methods and findings to empower science educators to make changes in their classrooms. My original contributions to knowledge can be found in the conceptual framework I developed (The Adaptive Critical Thinking (ACT) Framework); in reconfigured methodologies to gather and analyse my data; as well as in the novel findings about CT conceptions, development and assessment in Australia. My research findings have implications for how to conceptualise CT in Australia, and also inform how to obtain deeper insights into CT development.

Summary of key research findings and their implications

Through exploring the perspectives and actions of Australian science educators, this research generated a number of important findings. This section considers the methodological and theoretical contributions my research generates. The contributions are explored in relation to the aim they address, in the order they have been presented in this thesis. However, the contributions of the ACT Framework are described throughout this summary, because the use of this framework enhanced understanding of each of the classroom-based outcomes.

[27] Meta-analyses by Abrami et al. (2008, 2015) have shown that embedded approaches are the least effective method on their own, and educators should use a combination of explicit and embedded development opportunities to maximise CT development outcomes.

Findings and implications relating to Aim 1: To synthesise the theoretical construct of CT, to clarify the literature and generate an adaptable framework for CT and increase the accessibility of ideas on CT.

The first aim related to increasing accessibility of the theoretical construct of CT by clarifying understanding about the nature of CT. Synthesis and categorisation of the literature themes resulted in **a novel reconfiguration of CT theory into a framework which I have called the Adaptive Critical Thinking Framework** (Chapter 2). The ACT Framework and its elements help explain the CT process. They supply terminology for exploring CT literature (Chapter 1) as well as perceptions about CT (Chapter 3). This framework is accompanied by a definition to help guide its application.

The Adaptive Critical Thinking (ACT) Framework definition: *Critical thinking is a purposeful inquiry process that involves the deployment of thinking skills and context-appropriate criteria to make reasoned, reflective judgments and transform information into useable knowledge to resolve points of uncertainty.*

Importantly, this framework is adaptable to context. As outlined in Chapter 2, it can accommodate existing perspectives on CT, or provide a starting point to help educators to make decisions about how to present CT within their classroom. The framework can also be used in conjunction with other assessments to help identify aspects of CT missing from classrooms (Chapter 4 and 5). Further value of the framework to this research will also be described (in brief) in regards to aim 2 and 3 to show how this conception of CT informed the thinking in classrooms.

Findings and implications relating to Aim 2: To investigate Australian educator perceptions about CT and its development and assessment, and determine whether their understanding is consistent with Australian policy and/or the literature on CT.

To investigate Australian educator perspectives about CT, an adapted educator perception survey was used (containing ideas from education and ideas about CT) and results analysed using multivariate analysis techniques borrowed from ecology research (Chapter 3). This study involving the first cross-disciplinary, multi-institutional survey about CT in Australia resulted in two original contributions; it supplied new insights into Australian educator perspectives about CT; and an improved analysis method.

Studies conducted outside of Australia have previously identified weakness in educator understanding about CT (e.g. Paul et al., 1997a; Black, 2009; Choy and Cheah, 2009). The assumption underlying my perception study was that if Australian educators have gaps in their understanding, like their global peers, then the results of this research would highlight opportunities for targeted improvements that are contextually relevant for the Australian education system. However, if Australian educators were found to have a sound understanding of CT, then the policy-based approaches and teacher preparation methods used here could be treated as a model for improving educator understanding in other countries. Instead, knowledge addressing both scenarios has been generated.

While the study findings were new, in that educator perspectives in these contexts had not been researched before in Australia, the *a priori* findings from this first study were not overly surprising. **There were differences in high school and tertiary science and history educator perspectives about development and assessment. These differences were related to context-specific attributes (discipline-specific and education-setting).** However, it was also found that demography alone was not sufficient to clearly explain the conceptual differences in educator perspectives about CT. Suggesting there are things that influence an educator's values and ideas about CT that extend beyond an educator's immediate setting and background. Identifying these variables is outside the scope of this study. But it would be surprising if this difference is not partially a result of the range of perspectives about CT in the literature, and also the lack of a clear definition about CT in Australian education policy documents. However, it does highlight the importance of considering an educator's prior knowledge about CT to tailor training to meet educator needs and build on their existing understanding.

The second major contribution relating to this aim was an improved method for survey analysis. In fact, it was the **application of multivariate methods, the generation of global and sub-profiles of survey items and non-*a priori* analysis that facilitated the deepest insights into educator perceptions.** This global analysis of the survey results revealed 11 groupings that highlighted additional and unaccountable attributes that define the reasons why educators hold a certain perceptive about CT. This has implications for developing and running educator training about CT, as it is likely the audience will hold one of these 11 perspectives. However, future directions should build on this work to replicate these findings and develop a smaller set of questions that can be used to profile an educator's perspectives.

The findings from the education perception study increase understanding of Australian educator perspectives about CT from four contexts (high school educators, tertiary educators, science educators and history educators). Aside from the common theme of embedding CT approach in the classroom, **there was conceptual variety in Australian educator understanding of CT.** This variety in educator perspectives highlights the need for clearer guidance of the conception of CT through national education policies, because if educators are working from different understanding, then it will be challenging to offer resources to support them. However, the survey findings also revealed context-specific nuances about the way educators developed and assessed CT in science and history, and senior secondary and tertiary classrooms. This means any stance policy provides needs to be sensitive and adaptable to context. It signals the need for a conception of CT that can accommodate different perspectives (such as my ACT Framework). The use of the ACT Framework to explore educator responses about the concept of CT further clarified the differences in ideas about CT held by educators depending on their teaching experience, discipline and teaching environment (high school or university). It also supplied some initial insight into the conceptual perspectives driving the non-*a priori* statistically significant groupings identified in the global profile analysis.

Findings and implications relating to Aim 3: To explore how Australian tertiary science educators at Flinders University are developing CT and determine the effect of their approaches using a case-study method that could also serve as a general model for CT assessment.

To explore CT through the lens of development and assessment and generate understanding about actions and outcomes, two case studies were undertaken using observation, perception and performance measures. Individually, these studies increased understanding of CT development from an Australian first-year chemistry course, and an Australian science and society course. Collectively, these studies have contributed increased understanding about the measurable impact of the learning environment on CT development in science classrooms. These studies also supply an improved method for CT evaluation, demonstrating the capacity for a multi-tool classroom assessment process to help an educator diagnose and intervene in CT development.

Meta-analyses from the last ten or so years (such as Abrami et al., 2008, 2015) have suggested the smallest CT skill development outcomes result from embed CT classroom environments. Yet, the patterns in CT development changes observed in my case studies indicate that **the type of learning environment (active versus passive) rather than the inclusion of language around CT was the pedagogical component that led to the strongest CT performance outcomes**. These results have consequences for the courses used in these studies, as well as for tertiary science classrooms in Australia and the world. Importantly, they have shown that embedding CT in a discipline-specific course, the most common approach in Australian science classrooms (as revealed through Chapter 3), can lead to positive CT development outcomes. This is a particularly important finding for the Australian context, where there is an embedded stance in education policy documents at all levels (primary through to tertiary). However, not to be overlooked is the value of incorporating CT explicitly, because the explicit/infusion course had stronger indicators for perception and dispositional shifts than the embedded course (Chapter 5 and Chapter 6). While the specific role of affective factors in CT development are less understood, my results suggest the frame for CT development outcomes (being either skills-focused, disposition-focussed or a combination of the two) is important for determining which pedagogical changes an educator should incorporate to improve CT development in their context. However, further research using standardised disposition measures (such as the California Critical Thinking Disposition Inventory) is needed to clarify understanding about the more affective side of CT.

The secondary aim concerning the case study research was to develop and test the capacity of this case study method to serve as a general model for assessing and improving CT. CT assessment can also be approached in a number of ways (Lai, 2011). These range from using general to discipline-specific CT tests, and can further vary by being an explicit or embedded assessment approach. The findings from the case study explorations revealed opportunities to help Australian educators improve CT instruction in their classrooms. The case study components of my research demonstrated the capacity of a multi-tool approach by supplying better insights into CT development than any one instrument on its own. The classroom observations and perception survey results helped educators see beyond their intentions and unpack the collective experiences of their course, while the CAT supplied a quantitative performance measure. The addition of the ACT Framework to this reflection process helped identify aspects of CT missing from each course. While some issues with the CAT test and scoring alignment were encountered (discussed in Chapter 6), this instrument provided insights into student CT performance and generated discussion among educators that resulted in changes to future versions of the courses. Together the tools helped to discern the effect of a course and provided insights and options about ways to improve it.

The capacity for this approach to lead to positive outcomes in CT development was also tested in one of the courses. The information gleaned from the initial study was used to make evidence-informed changes. These included pedagogical changes to the course coordinators delivery, increased scaffolding of the process of thinking critically, increased discussion involving CT terminology, as well as the incorporation of CAT-like CT intervention tasks to increase the amount of individual problem-solving. These changes were found to have a positive effect on critical thinking performance, with significantly stronger performance outcomes than measured in the previous version of the course. This demonstrated the capacity for using this case-study method to improve student CT skills, suggesting my multi-tool assessment approach is useful for identifying opportunities to make specific changes in a course. Additionally, this approach showed that improved CT development outcomes could be achieved without making extensive changes to a course. While a limitation of these studies is they were all conducted at one institution, the implications for this case-study process are not institution-specific. The assessment tools have been individually used in many classrooms around the globe, and the course-level analyses undertaken require only basic statistical knowledge.[28] Therefore, this method has great potential for providing educators with a means to evaluate and improve CT development in their classroom. However, further studies using this approach would need to be undertaken to verify the degree of success of the method in other contexts, and also help educators determine the minimum amount of changes needed to produce a stronger CT outcome.

The value of transdisciplinary methods

Through this transdisciplinary journey I considered whether the current understanding (knowledge and methods) in each space (conception, development and assessment of CT) was sufficient to support educators in their endeavour to develop the next generation of thinkers. However, existing theories and approaches were often found to be inaccessible to educators or inappropriate for their contexts. Further, while individual studies about CT have found differences in CT performance across different learning settings, much of the research has been unable to address key questions about the specific contributions of the classroom experiences to CT development. This body of work has explored classroom-based perspectives through the lense of CT to provide evidence-informed insights into CT conceptions, development and assessment.

[28] This process is simplified for educators because the CAT developers supply a report for their test results upon request. However, I completed my own analyses, which were slightly more complex but more statistically appropriate. But even with the more complicated analyses, the general trends can be discerned from the bar graphs, with the test statistics providing verification of the visually discernable differences.

By the very nature of the role, teachers in higher education already work at the face of two intersecting domains – their discipline (such as science or history) and education. As Schulman (2000) points out, there is a responsibility to engage in the scholarship of one's professional field, and one's profession as an educator. However, in science, there tends to be more emphasis on scholarship related to one's professional field. Of those science educators who have forged ahead and examine their teaching practice, it has been suggested that the way science educators undertake their work puts them at risk of fragmentation and isolation from other fields (Talanquer, 2014; Henderson, Beach and Finkelstein, 2011). Recent findings by Trujillo and Long (2018) support these claims, as their study on co-citation analysis in systems thinking revealed that the way STEM education and DBER is undertaken demonstrates weak connections to related knowledge communities from other fields. To avoid isolation and to strengthen the scholarship of science education research, it has been suggested that science educators need to borrow concepts and methodologies from other fields (Singer et al., 2012; Dolan, 2015; Wooten, 2019). Yet, this dissertation has shown the power of making this a two-directional relationship.

By applying my biology-based science training to my doctoral research, I was able to enhance the conceptual and pedagogical understanding gained from all my research components. For example, the ACT Framework arose out of applying a taxonomy-like classification approach to CT theory, where I look for shared characteristics across the different perspectives presented in the literature (Chapter 2). Ecologically-based multivariate analysis approaches were applied in my educator perception study (Chapter 3). This resulted in the identification of conceptual relationships across educator groups that were not evident using traditional survey analysis approaches. Importantly this approach also bridges the gap between applying univariate statistics to small-scale survey studies and regression analyses which rely on having a large (and ideally normally distributed) data set. Finally, my case studies reconfigured existing education assessment resources into a multi-tool approach to generate perception, performance and observation data. While classroom environments present challenges scientists usually would not face (for example, students cannot be made to retake a test, or exposed to trial after trial after trial), biologically based thinking still informed how I undertook these classroom explorations. Classroom environments are not that dissimilar from doing field-based observational studies, where many factors cannot be controlled or manipulated. However, a key difference between typical classroom-based studies and biology-based observation studies is that a scientist would never take one kind of measurement in an uncontrolled environment to explain an effect in that environment. This meant a multiple-measures approach with controlled conditions was needed. The application of a multi-tool approach in my studies made it possible to generate deeper insights into CT development in classroom environments and effects (Chapter 4, 5 and 6), resulting in pedagogical knowledge that could facilitate significant change in performance outcomes from one year to the next (Chapter 5).

Future directions for CT development and assessment

There are ongoing calls for increased research into 21st-century skills (Arum et al., 2016; Doeke and Maire, 2019). In this rapidly changing age of knowledge, we need science educators to be thinking about their classrooms as an opportunity to increase their understanding about student learning, and as an opportunity for students to gain transferrable skills. This requires STEM educators to balance multiple goals and knowledge. They need epistemological content knowledge concerning the discipline they teach. They need pedagogical knowledge to make informed decisions about the way they will deliver opportunities to foster student learning. They also need knowledge of CT and the subtleties of its application in their discipline so they can determine which aspects to target with students and at what time. These challenges that educators face are not only true for CT, but also for making other pedagogical changes (Henderson et al., 2011; Brownell and Tanner, 2012; Carbone et al., 2019). However, it is particularly problematic for decisions relating to CT because previous studies indicate educators do not have the time nor the training to evaluate which approaches and resources to adopt in their practice (Paul et al., 1997a; Black, 2009; Alwadai, 2014). While the Australian educators surveyed in the educator perception survey study (see Chapter 3) were not explicitly asked about professional development relating to CT, it was evident through their overall responses about CT that there were varying levels of understanding and experiences among respondents.

Science educators must be as methodological and accountable to their teaching goals/outcomes as they would be to undertake a science experiment to further knowledge in their field. They need to gain feedback from their classrooms so they can understand what needs to change. Shipley et al. notes that to achieve improvement in STEM education, "a cycle of applying research in context and drawing from education practice to create and refine powerful learning theories" is needed (2017, p.358). Changes made in first year provide seeds of development for future years. It also provides the possibility for multiple opportunities to practice CT not just within a course, but across a degree. Shulman (2000) explains that while education research can require considerable effort by an educator, the payoff is a sounder education system – one where student, institutions, one's profession and policy all benefit from a higher standard of education. However, not all the responsibility lies on the individual. Education systems, particularly higher education, shape the decision educators make through the metrics they use for job evaluation and progression. So, while it is not uncommon for university science teachers to be researchers in their classrooms (Abell, 2005), research in one's professional field tends to be prioritised over the scholarship of teaching research due to the criteria for job progression.

Despite updates to Australian education policies over the last ten years, CT and other general capabilities remain poorly integrated into Australian classrooms (Doeke and Maire, 2019). This is likely a product of the fact that they are under-described in policy documents. A result of this lack of clear direction about development, they are under-assessed by educators who do manage to embed CT in some way. In Australia, assessment of CT is generally embedded within content, and as a result, CT evaluation becomes a small fraction of the overall grade. This embedding and mark allocation has a number of consequences. Firstly, it is not always apparent to students that they are doing CT, especially if the assessment terminology is different from the language used in class. Secondly, embedding CT among other learning outcomes could also mean it is not seen as important as content by students because it is not given many marks (Choy and Cheah, 2009). Additionally, it is harder to interpret and explain the difference between no CT, developing CT and high-quality CT, if there are only a few marks to allocate. Having poorly outlined goals and expectations also makes it harder for students to be successful (Roksa et al., 2017). So there are numerous reasons why educators need to be mindful and specific about where and how CT is included in their classrooms, even if the delivery of the CT associated activities reflects an embedded CT approach.

Educators must take care to align the theory, tasks and scope when designing their own CT assessments. Carefully designed assessments are important to foster increase academic rigour and produce stronger intellectual development (Culver et al., 2019). However, when educators are left to create their own CT assessments, it can result in ill-defined criteria (Elliot et al., 2010a). Self-developed approaches, the preference for the majority of educators for most assessment items in my perception survey (Chapter 3), rarely allow for scrutiny, validation and replication studies. The consequence of this is the educator may be using a tool that inadvertently overstates the strength of their practice. Given the general lack of expertise and training about CT skills assessments among educators, most educators would benefit from clearer policy-based direction about CT assessment including a structured approach. Policymakers could recommend using either a general or discipline-specific standardised CT assessment. Yet this advice should come with a proviso - as just using any standardised measure does not guarantee accurate findings for any particular classroom. Hatcher's (2013) varied findings from the same course when measured with different instruments are a clear example of this. Standardised assessments have been criticised for not capturing "what goes on in the classroom between the teacher and the student" (Benjamin, 2012, p.10). However, Benjamin (2012) also points out that this does not override the need to assess what is going on; it just means that assessment should include multiple methods that "can capture… the complexity [of courses, programs and colleges]" (Benjamin, 2012, p.10). This recommendation echoes the Delphi panel's recommendation to use multiple instruments to capture a holistic view of CT (Facione, 1990b). Benjamin (2012) has suggested the need for tools that can provide educators with information that facilitates understanding of the CT development outcomes from their course, and what they might target to achieve best practice. A general approach, such as my case-study method (Chapter 4), would allow educators to make informed judgements when designing programs and choosing assessments. My approach diagnosed the classroom environment, student performance and perceptions, with the information gained subsequently used to modify pedagogy and improve CT outcomes. While I have found valuable insights from this set of tools, the specific assessment instruments used could be swapped to equivalent types that better suit the context and/or the current pedagogical focus of the educator (such as an essay-based CT test, a multiple-choice CT test, and/or a standardised disposition-based CT test). But until otherwise required through policy, it is ultimately up to educator discretion. Educators need to be satisfied they have thoughtfully assessed the CT development opportunities they are providing for their students and are not just readily accepting a particular type of feedback because it is easy to gather or reinforces their beliefs.

In a US-based study, Leming (2016) noted that educators' choices regarding instruction and assessment were connected to their prior education experiences of approaches that had helped them learn. Similarly, Bathgate et al. (2019b) found that educator choices are influenced by personal opinion and experience. However, Bathgate et al. (2019a, 2019b) also noted the role of workplace culture in shaping an educator's decisions. Importantly, Bathgate et al. (2019b) and Shadle et al. (2017) found that a supportive work environment (such as access to resources, and support from colleagues) increased the likelihood of an educator making evidence-based changes to their practice. Therefore, there also needs to be increased educator training in CT theory and development. However, as found in my educator perception survey, there are differences in educator perspectives as well as context-specific ways that CT is approached. To create a more cohesive and universal approach, the policymakers in the Australian education system should identify the CT conception they want Australian educators to incorporate into their practice, and should offer relevant training and resources in line with this perspective. However, some caution should be taken with a more globalised approach because “real-world assessment problems resist one-size-fits-all solutions” (Wright et al., 2017, p.45). This is where the case-study method and the ACT Framework presented in my research offer solutions. The ACT Framework provides a broad overview of the types of things an educator should consider when designing for CT, while the case-study method provides a general approach that will supply insights into the classroom environment and its CT outcomes. However, additional studies need to be undertaken to help generate a collective dataset that will be useful for determining best-practice, and also for verifying the applicability of this method in new contexts.

Final thoughts

Abell notes there are a “variety of purposes for engaging in university teacher research”, including “the need to understand and improve our teaching,” “generating knowledge for audiences beyond one’s classroom,” in addition to building “relationships across academic units” such as education and science (Abell, 2005, p.292). It is not necessary to be motivated by every purpose, but most apply to the inspiration behind my doctoral journey. All the components of this research sought to generate knowledge and understanding that would help facilitate direct impacts on CT development for individual educators at a classroom level. This began with a synthesis of the theoretical construct of CT, followed by explorations of educator perspectives and CT development in science classrooms, and resulted in the creation of an adaptable framework for CT designed to increase the accessibility and translatability of ideas related to CT for educators. Further through the use of a novel combination of data gathering tools and innovative applications analysis techniques used in my studies, this research has provided some new directions for exploring CT development and for conducting multivariate research in education (and for applying multivariate analysis techniques to survey and classroom-based data sets to bring deeper insights).

This dissertation has shown that the journey CT takes from concept to classroom is complex. It has explored CT across different theoretical fields. It moves knowledge across boundaries, borrowing methodologies and insights from science, science education and curriculum development to refine understanding about CT conceptions, CT development and CT assessment. This transdisciplinary research has considered these three lenses of CT from two different angles, looking at existing knowledge and existing methods before reconfiguring these aspects to produce more profound insights into CT. This led to original insights into methods used to explore and approach CT development, as well as generating new insights about CT in specific classrooms. Further, through these classroom insights greater understanding about the nature of CT development has been gleaned, specifically that active learning rather than passive learning results in greater CT scores. Collectively my chapters have demonstrated the interplay between concepts informing classrooms and classrooms informing concepts.

This research is a product of both my own experiences and a synthesis of knowledge gained from experts and stakeholders. I have been mentored by numerous science educators whose experiences have covered the full Australian education system as well as some perspectives from around the globe. I have also had the benefit of working with three amazing psychology postdocs, a political scientist, a cultural studies professor and a statistician. All these relationships have functioned as soundboards, and have pushed me to think about things from different perspectives. This journey has led me to shift from viewing connections between knowledge, which Boon and Van Baalen (2019) describe as a network of theories, to understanding the power of the movement of knowledge across disciplinary boundaries.

This thesis has shown the power of transdisciplinary research approaches for uniting science education research with other connected fields and applicable knowledge. An essential aspect of transdisciplinary methods is the movement of ideas (Somerville and Rapport, 2000). Through this research, I bring ideas from science, critical thinking and curriculum development to create new understanding about CT development. However, the understanding generated also feeds back into the disciplines I borrowed ideas from. While this research has a definite emphasis on the exploring way things are done in Australia, there are some more mobile components of my research. The lessons from CT development apply to global curriculum design in science and add to understanding about the effects of embedded approaches (relevant to Australia outside of science). However, the framework and its associated definition are the most mobile component of my research. They can be used to delve into CT theory - as a start place for thinking about CT, and/or they can be used as an information processing tool to navigate and filter the literature on CT. They can be used to guide learning. For example, students can use the framework to self-reflect on their thinking. They can be used to guide practice and assessment. For example, educators can use the framework to make formative evaluations of student work, and/or reflect on their practice to help recognise opportunities to target specific aspects of CT.

As I reach the end of this journey, at least for the context of this dissertation, I have come to recognise that not only does transdisciplinary research need CT, but CT itself is also transdisciplinary. CT is a mobile construct that is nuanced in its application. This is perhaps why it has been so hard for theorists to generate a consensus stance for CT. A reductive CT model will always encounter a scenario where the rules do not fit. This will lead to claims that consensus has not been achieved and renewed calls for a common definition (as has occurred multiple times in the last 70 or so years). This is not an unreasonable process, as theories are often updated in line with new understanding – especially in the field of science. But is it rational? Perhaps not - which is why the biggest irony of this journey is the production of a potential new consensus model through identifying why consensus has still not been achieved. My critical thinking framework and associated definition are the product of the reconfiguration of existing ideas, rather than creating something entirely new. I focussed on constructing a general approach that would transcend any single context, but could be readily adapted to capture the discipline-specific nuances and knowledge. Yet, while ironic, it should not have been surprising, as knowledge generation is the very essence of a transdisciplinary nomadic path.

As with all doctoral candidates who have gone before me - imposter syndrome chips away at my confidence, leaving me feeling like a fish out of water. Yet, I wonder, who else could have reconfigured knowledge about CT in this way? My configuration of qualifications and teaching experience are unique. My mentors challenged me to challenge conventional ways of knowing. So, as I conclude this doctoral journey, I invite challenges to this reconfigured perspective, because as was stated to me by Professor Westwell at the beginning of this research process - the best way to figure out what CT is, is to consider what it is not.

APPENDICES

APPENDIX A Educator perception study survey template

Q1) Please allocate points to the following statements based on how well they fit your beliefs about CT. You have a total of 15 points to allocate. You can choose one statement (at 15 points) or spread your points across multiple statements. The statement most relevant to your beliefs about critical thinking should receive the highest point allocation, and irrelevant statements can receive minimal or no point allocation. *(Point spending question format)*

Note: If using this survey question and you are not trained in multivariate data analysis, you may wish to get educators to rank the statements instead of distributing points.

Statement	Points
Critical thinking requires a combination of dispositions, knowledge, and skills, all of which develop over time.	
General critical thinking skills are readily transferable to discipline-specific contexts without additional scaffolding.	
Critical thinking skills should be developed through approaches which are embedded across the curriculum.	
Critical thinking requires both thinking skills and dispositions, which are mapped against standards.	
Critical thinking skills should be developed explicitly and separately from other content matter.	
Critical thinking consists of higher-order thinking skills such as analysis, synthesis and evaluation.	
Critical thinking is embodied by step by step procedures and skill development leading to higher levels of thinking.	
To be effective critical thinkers, students require specific thinking skills and knowledge related to their discipline of study.	
Critical thinking skills require scaffolding to be transferable into new contexts.	
Standards and criteria are essential elements to critical thinking.	
Critical thinking is primarily concerned with the application of the formal rules of logic.	

Q1a) What features, skills, traits and dispositions do you associate with critical thinking? Please list as many as you can think of. *(Open-ended extended response question format)*

Q2) For each of the following competencies and capabilities listed below, indicate the level of importance when compared to critical thinking. *(Likert scale question format)*

	Options				
Learning competency	Learning competency much more important compared to critical thinking	Learning competency more important compared to critical thinking	Learning competency and critical thinking have equal importance	Learning competency less important compared to critical thinking	Learning competency much less important compared to critical thinking
Creativity & innovation					
Knowledge					
Analysis					
Teamwork					
Reasoning					
Intercultural understanding					
Understanding (content)					
Numeracy					
Communication					
Information & communication literacy					
Literacy					
Citizenship					
Planning					
Problem Solving					
Ethical conduct & understanding					

Q3) Do you intentionally set out to develop critical thinking skills in your students? *(Pick an option question format)*

Pick one option	Choice	Follow up question
Yes with approaches involving explicit instruction		See 3a
Yes with approaches embedded across the curriculum		See 3b
No		See 3c

3a) Series of three questions as denoted by i-iii

i) Please provide an example of the program or approach you use to develop your student's critical thinking skills. *(Open-ended extended response question format)*

ii) Have you considered assessing this teaching approach to determine its effectiveness for developing critical thinking skills and abilities? *(Pick an option question format)*

Option	Choice
Yes but have only considered it	
Yes – have both considered and assessed it	
No	

iii) Do you have any further comments about this? *(Open-ended question format)*

3b) How do you ensure that critical thinking is developed in your students by this embedded approach? *(Open-ended extended response question format)*

3c) Please explain how you know if critical thinking skills are an emergent property of your students' learning? *(Open-ended extended response question format)*

Q4) Which (if any) of these approaches have you found to effective for developing critical thinking abilities in your students? *(Likert scale question format)*

Teaching approach	Options					
	Much more effective for developing critical thinking abilities	More effective for developing critical thinking abilities	Neither effective nor ineffective for developing critical thinking abilities	Less effective for developing critical thinking abilities	Not effective for developing critical thinking abilities	Have not tried*
Argument analysis						
Constructing critiques						
Writing tasks						
Teamwork						
Problem-based learning						
Flipped classroom						
The scientific method						
Oral presentations						
Context- dependent sets						
Peer assessment						
Collaborative learning						
Concept mapping						
Debating						
Case studies						
Socratic questioning						
Deductive reasoning activities						
Inductive reasoning activities						
Logic modes						
Other~						

~ For this option educators were also given the opportunity to specify their own teaching-related activity and indicate the effectiveness of the activity for developing critical thinking.
* This option was originally described as 'Unsure/ have not tried.' However, these mean different things, so for more meaningful results when using this question, it is suggested that you only include this as 'have not tried'.

Q5) Which of the following strategies do you think you would use to assess critical thinking at your educational institution? *(Pick an option question format)*

Assessment tool type	Options			
	Self-developed item	Published resource*	Purchased resource*	Would not use
Course evaluation form				
Discipline-specific critical thinking skills test				
Extended response items				
General critical thinking skills test				
Multiple ranking items				
Multiple-choice items				
Rubric				
Self-report items				

*Note: Respondents did not seem to distinguish between these two items, so they were combined for analysis in this study. However, the two categories were intended to separate assessment tools which were 'published in a peer-reviewed article – i.e. evidence-based' versus those tools which were 'an externally purchased resource that may not be published or evidence-based.'

Q6) Allocate points to the following statements based on how well they describe your current approach to assessing critical thinking. You have 5 points to allocate. Indicate your choices by allocating points to one or more statements. The most relevant statement to your approach should receive the highest point allocation and irrelevant statements can receive minimal or no point allocation. *(Point allocation question format)*

Note: If using this survey question and you are not trained in multivariate data analysis, you may wish to get educators to rank the statements instead of distributing points.

Statement	Points
I have difficulty identifying student work that reflects critical thinking.	
I do not look for specific evidence of critical thinking when assessing students work.	
I do not look for specific evidence of critical thinking when assessing students work, but know my students are developing this skill as their marks improve.	
I look for specific evidence of critical thinking embedded in students work.	
I assess critical thinking development using self-report assessment items.	
I assess critical thinking development using multiple-choice items assessment items.	
I use rubrics to assess critical thinking in written responses.	
I use a combination of rubrics, self-report and multiple-choice items to assess critical thinking.	
I assess critical thinking development using published critical thinking test items.	

6a) Are there any other strategies you would implement to assess critical thinking development? *(Open-ended extended response question format)*

Demographic information requested

Gender:

State/ Territory:

Discipline:

Type of educational institution: (i.e. High school or tertiary)

How many years have you been teaching? (This data was grouped from 1-5years; 6-15years; 15+ years)

APPENDIX B Educator perception study survey keys

Table B.1 Australian senior secondary and university educator perception survey key for question set relating to the global profile. Statement choices listed in corresponding sub-profiles.

Global profile		
Question	**item codes**	**Corresponding sub-profile**
Q1) Please allocate points to the following statements based on how well they fit your beliefs about CT.	C_1a - C_1k	Conceptual (11 items)
Q2) For each of the following competencies and capabilities listed below, indicate the level of importance when compared to CT.	D_2a - D_2o	Developmental (36 items)
Q3) Do you intentionally set out to develop CT skills in your students?	D_3a – D_3c	
Q4) Which (if any) of these approaches have you found to effective for developing CT abilities in your students?	D_4a – D_4r	
Q5) Which of the following strategies do you think you would use to assess CT at your educational institution?	A_5a – A_5h	Assessment (33 items)
Q6) Allocate points to the following statements based on how well they describe your current approach to assessing CT.	A_6a – A_6i	

Table B.2 Australian senior secondary and university educator perception survey key for questions and statements relating to the conceptual profile.

Conceptual sub-profile	
Question 1 (Q1): Please allocate points to the following statements based on how well they fit your beliefs about CT. You have a total of 15 points to allocate. You can choose one statement (at 15 points) or spread your points across multiple statements. The statement most relevant to your beliefs about CT should receive the highest point allocation, and irrelevant statements can receive minimal or no point allocation.	
KEY	**Statement choices (11 items)**
C_1a	CT requires both thinking skills and dispositions, which are mapped against standards.
C_1b	CT requires a combination of dispositions, knowledge, and skills, all of which develop over time.
C_1c	CT consists of higher-order thinking skills such as analysis, synthesis and evaluation.
C_1d	CT is embodied by step-by-step procedures and skill development leading to higher levels of thinking.
C_1e	Standards and criteria are essential elements to CT.
C_1f	To be effective critical thinkers, students require specific thinking skills and knowledge related to their discipline of study.
C_1g	CT skills should be developed through approaches which are embedded across the curriculum.
C_1h	CT skills should be developed explicitly and separately from other content matter.
C_1i	CT skills require scaffolding to be transferable into new contexts.
C_1j	General CT skills are readily transferable to discipline-specific contexts without additional scaffolding.
C_1k	CT is primarily concerned with the application of the formal rules of logic.

Table B.3 Australian senior secondary and university educator perception survey key for questions and statements relating to the CT development.

Developmental sub-profile	
Question-set 2(Q2): For each of the following competencies and capabilities listed below, indicate the level of importance when compared to CT. (Likert scale of importance)	
KEY	**Attributes for comparison (15 items)**
D_2a	analysis
D_2b	citizenship
D_2c	communication
D_2d	creativity
D_2e	ethical conduct
D_2f	ICT
D_2g	intercultural understanding
D_2h	knowledge
D_2i	literacy
D_2j	numeracy
D_2k	planning
D_2l	problem-solving
D_2m	reasoning
D_2n	teamwork
D_2o	understanding
Question 3 (Q4): Do you intentionally set out to develop CT skills in your students?	
KEY	**Statement options (3 items)**
D_3a	Yes - with approaches involving explicit instruction
D_3b	Yes - with approaches embedded across the curriculum
D_3c	No plan
Question-set 4 (Q4): Which (if any) of these approaches have you found to effective for developing CT abilities in your students? (Likert scale on effectiveness)	
KEY	**Teaching approach options (18 items)**
D_4a	problem-based learning
D_4b	collaborative learning
D_4c	concept mapping
D_4d	logic models
D_4e	context-dependent sets
D_4f	debating
D_4g	flipped classroom
D_4h	scientific method
D_4i	Socratic questioning
D_4j	peer assessment
D_4k	inductive reasoning activities
D_4l	deductive reasoning
D_4m	constructing critiques
D_4n	class discussion
D_4o	oral presentation
D_4p	case studies
D_4q	writing tasks
D_4r	argument analysis

Table B.4 Australian senior secondary and university educator perception survey key for questions and statements relating to the CT assessment.

Assessment sub-profile	
Question-set 5 (Q5): Which of the following strategies do you think you would use to assess CT at your educational institution?	
KEY	**Assessment tool options (24 items)**
A_5a - SD	Course evaluation form - Self - developed item
A_5a - ER	Course evaluation form - External resource
A_5a - NO	Course evaluation form - Would not use
A_5b - SD	Discipline-specific CT skills test - Self - developed item
A_5b - ER	Discipline-specific CT skills test - External resource
A_5b - NO	Discipline-specific CT skills test - Would not use
A_5c - SD	Extended response items - Self - developed item
A_5c - ER	Extended response items - External resource
A_5c - NO	Extended response items - Would not use
A_5d - SD	General CT skills test - Self - developed item
A_5d - ER	General CT skills test - External resource
A_5d - NO	General CT skills test - Would not use
A_5e - SD	Multiple ranking items - Self - developed item
A_5e - ER	Multiple ranking items - External resource
A_5e - NO	Multiple ranking items - Would not use
A_5f - SD	Multiple-choice items - Self - developed item
A_5f - ER	Multiple-choice items – External resource
A_5f - NO	Multiple-choice items - Would not use
A_5g - SD	Rubric - Self - developed item
A_5g - ER	Rubric - External resource
A_5g - NO	Rubric - Would not use
A_5h - SD	Self - report items - Self - developed item
A_5h - ER	Self - report items - External resource
A_5h - NO	Self - report items - Would not use
Question-set 6 (Q6): Allocate points to the following statements based on how well they describe your current approach to assessing CT. You have 5 points to allocate. Indicate your choices by allocating points to one or more statements. The most relevant statement to your approach should receive the highest point allocation and irrelevant statements can receive minimal or no point allocation.	
KEY	**Statement options (9 items)**
A_6a	I have difficulty identifying student work that reflects CT.
A_6b	I do not look for specific evidence of CT when assessing students work.
A_6c	I do not look for specific evidence of CT when assessing students work, but know my students are developing this skill as their marks improve.
A_6d	I look for specific evidence of CT embedded in students work.
A_6e	I assess CT development using self-report assessment items.
A_6f	I assess CT development using multiple-choice items assessment items.
A_6g	I use rubrics to assess CT in written responses.
A_6h	I use a combination of rubrics, self-report and multiple-choice items to assess CT.
A_6i	I assess CT development using published CT test items.

APPENDIX C Additional results from Chapter 3: The Australian educator perception survey

Table C.1 Summary of findings for *a priori* hypothesis tests undertaken within individual survey question-sets, using PERMANOVA to explore if the global and sub-profiles masked the effects of key demographic factors on educator perceptions about critical thinking. Symbols added to indicate significant differences between the 0.01 and 0.05 significance level.

	conceptual sub-profile			development sub-profile									assessment sub-profile					
	Q1			Q2			Q3			Q4			Q5			Q6		
	conceptual perspectives			attributes and outcomes			development approach			classroom perspectives			assessment tools			assessment perspectives		
factor	pseudo-F	p-value	df	pseudo-F	p-value	df	pseudo-F	p-value	df	pseudo-F	p-value	df	pseudo-F	p-value	df	pseudo-F	p-value	df
discipline	0.37	0.8442	1	2.25	0.0479	1	1.27	0.2747	1	**8.94**	**0.0001**	**1**	1.95	0.1014	1	2.97	0.0159`	1
education setting	0.08	0.9689	1	**4.98**	**0.0002**	**1**	0.21	0.7694	1	**4.12**	**0.0003**	**1**	**9.74**	**0.0001**	**1**	1.18	0.3087	1
gender	2.97	0.0121`	1	1.20	0.3093	1	0.21	0.7880	1	2.04	0.0369`	1	1.87	0.1154	1	1.83	0.1115	1
state	1.39	0.0915	6	0.70	0.8593	6	0.64	0.7655	6	1.01	0.4266	6	1.32	0.1430	6	1.00	0.4534	6
teaching experience	0.18	0.9888	2	1.32	0.2068	2	0.39	0.7675	2	1.51	0.0929	2	1.92	0.0592	2	1.06	0.3882	2
discipline vs education setting	0.74	0.5963	1	0.00	>0.05	1	0.37	0.6297	1	0.84	0.5616	1	1.45	0.2203	1	0.13	0.9669	1
discipline vs gender	1.18	0.3332	1	0.89	0.5062	1	1.14	0.2865	1	0.80	0.6116	1	1.76	0.1508	1	0.10	0.9719	1
discipline vs state*	0.97	0.4980	4	0.61	0.8643	4	0.63	0.7184	4	0.83	0.6226	4	1.26	0.2145	4	0.72	0.7904	4
discipline vs teaching experience	0.48	0.8862	2	0.56	0.8251	2	0.25	0.8747	2	0.96	0.4534	2	0.74	0.6701	2	1.60	0.1048	2
education setting vs gender	0.88	0.4940	1	0.89	0.4972	1	0.08	0.8971	1	1.26	0.2672	1	0.30	0.8482	1	0.67	0.6337	1
education setting vs state*	0.78	0.7217	4	1.56	0.0645	4	0.89	0.5090	4	1.40	0.0915	4	1.34	0.1750	4	1.31	0.1757	4
education setting vs teaching experience	0.62	0.7841	2	0.53	0.8676	2	0.43	0.7348	2	1.46	0.1194	2	0.90	0.5337	2	1.84	0.0604	2
gender vs state*	1.54	0.0431`[#]	6	1.25	0.2156	6	1.00	0.4171	6	1.34	0.1211	6	0.76	0.7165	6	1.04	0.4110	6
gender vs teaching experience	2.37	0.0122`~	2	1.12	0.3454	2	0.84	0.4635	2	0.82	0.6550	2	1.05	0.4035	2	0.30	0.9721	2
state* vs teaching experience	1.32	0.1062	8	0.81	0.7576	8	0.86	0.5847	8	1.18	0.2040	8	1.00	0.4705	8	0.99	0.4907	8
science - education setting	0.37	0.8530	1	**4.16**	**0.0013**	**1**	0.21	0.7641	1	**4.10**	**0.0002**	**1**	**8.63**	**0.0001**	**1**	1.03	0.3877	1
science - gender	2.88	0.0148`	1	0.54	0.7608	1	0.87	0.3676	1	2.18	0.0232`	1	1.55	0.1998	1	1.30	0.2578	1
science - state	1.39	0.0850	6	0.69	0.8687	6	0.79	0.6426	6	0.95	0.5163	6	1.52	0.0531	6	1.11	0.3114	6
science - teaching experience	0.11	0.9964	2	0.92	0.5221	2	0.17	0.9589	2	1.47	0.1039	2	1.39	0.2053	2	0.86	0.5620	2

*pairwise tests incorporating state as a factor excluded NT and ACT respondents from analysis due to low sample size.

`pairwise tests for state at the global level revealed four important differences: Vic HS vs Vic TER (t= 1.87 p=0.0001); NSW HS vs NSW TER (t= 2.42, p=0.0001); SA HS vs Vic HS (t=1.55 , p=0.0076); SA HS vs WA HS (t= 1.42, p=0.0352)

[#]pairwise tests for gender-state at the conceptual level identified seven key differences: male VIC vs female QLD (t=1.78, p=0.0127); female VIC vs female QLD (t=1.77, p=0.0139); male VIC vs male WA (t=1.70, p=0.0218); female QLD vs female NSW (t=1.66, p=0.0372); male NSW vs female NSW (t=1.57, p=0.0407); male VIC, male QLD (t=1.53, p=0.0414); female ACT vs female VIC (t=1.61, p=0.0425).

~pairwise tests for gender - teaching experience at the conceptual level revealed 2 differences: male 6-15years vs male 1-5years experience (t=1.75, p=0.0122); male 1-5years experience vs female 1-5years experience (t=1.74, p=0.217); male 6-15years experience vs female 6-15years experience (t=1.65, p=0.0278).

Table C.2 Summary of findings for *a priori* hypothesis tests undertaken on the global and sub-profiles, using PERMANOVA to explore if any particular demographic factors underpinned educator perceptions about critical thinking. Symbols added to indicate significant differences between the 0.01 and 0.05 significance level.

	global profile			conceptual sub-profile			developmental sub-profile			assessment sub-profile		
factor	pseudo-F	p-value	df	pseudo-F	p-value	df	pseudo-F	p-value	df	pseudo-F	p-value	df
discipline	**4.25**	**0.0001**	**1**				**7.14**	**0.0001**	**1**	2.03	0.0776	1
education setting	**7.45**	**0.0001**	**1**				**6.02**	**0.0002**	**1**	**8.98**	**0.0001**	**1**
gender	1.77	0.0645	1				1.15	0.3109	1	1.86	0.1112	1
state	1.11	0.2677	6				0.89	0.6432	6	1.29	0.1512	6
teaching experience	1.39	0.1193	2				0.97	0.4690	2	1.85	0.0603	2
discipline vs education setting	0.68	0.7444	1				0.46	0.8653	1	1.08	0.3794	1
discipline vs gender	1.12	0.3416	1				0.76	0.6177	1	1.62	0.1607	1
discipline vs state*	1.08	0.3345	4	The set of items in this question are the same as the conceptual profile item set. Values for this are displayed in Table C.1			0.72	0.8414	4	1.21	0.2524	4
discipline vs teaching experience	0.87	0.6165	2				0.85	0.5998	2	0.83	0.5986	2
education setting vs gender	0.52	0.8782	1				0.91	0.4981	1	0.32	0.8698	1
education setting vs state*^	1.63	0.0106'^	**4**				1.42	0.0810	4	1.40	0.1260	4
education setting vs teaching experience	0.72	0.8054	2				0.58	0.8726	2	0.97	0.4630	2
gender vs state*	0.95	0.5717	6				1.37	0.0756	6	0.64	0.9203	6
gender vs teaching experience~	1.07	0.3768	2				0.79	0.6700	2	0.93	0.5088	2
state* vs teaching experience	0.95	0.6044	8				0.72	0.9234	8	1.01	0.4615	8
science: education setting	**6.41**	**0.0001**	**1**				**5.36**	**0.0003**	**1**	**7.89**	**0.0001**	**1**
science: gender	1.49	0.1326	1				1.22	0.2749	1	1.55	0.1784	1
science: state	1.19	0.1554	6				0.81	0.7857	6	1.48	0.0548	6
science: teaching experience	1.14	0.3031	2				0.87	0.5787	2	1.35	0.2115	2

*pairwise tests incorporating state as a factor excluded NT and ACT respondants from analysis due to low sample size.

` given the exploratory nature of this study, these factors and interactions were further explored even though they exceeded the 0.01 cut off.

^pairwise tests for state at the global level revealed four important differences: VIC HS vs VIC TER (t= 1.8547 p=0.0002); NSW HS vs NSW TER (t= 2.4059, p=0.0001); SA HS vs VIC HS (t=1.5471 , p=0.008); SA HS vs WA HS (t= 1.4212, p=0.0366).

Table C.3 Proportions of educators who indicated agreement with corresponding Likert-options for Question 2. Displayed by key factors: discipline, and education setting (everyone and within the science-only cohort).

			Proportions by demographic factors#					
			Discipline		Education setting		Science: Education setting	
Item	Item description	Options	Science (n=110)	History (n=36)	High school (n=96)	Tertiary (n=50)	High school (n=69)	Tertiary (n=41)
D_2a	analysis	CT less important (1 -2)	25%	17%	24%	20%	26%	22%
		equal importance (3)	71%	81%	73%	74%	71%	71%
		CT more important (4-5)	5%	3%	3%	6%	3%	7%
D_2b	citizenship	CT less important (1 -2)	7%	6%	9%	2%	10%	2%
		equal importance (3)	35%	33%	41%	22%	42%	22%
		CT more important (4-5)	58%	61%	50%	76%	48%	76%
D_2c	communication	CT less important (1 -2)	33%	28%	40%	16%	43%	15%
		equal importance (3)	61%	67%	59%	68%	57%	68%
		CT more important (4-5)	6%	6%	1%	16%	0%	17%
D_2d	creativity	CT less important (1 -2)	21%	19%	25%	12%	28%	10%
		equal importance (3)	63%	67%	60%	70%	57%	73%
		CT more important (4-5)	16%	14%	15%	18%	16%	17%
D_2e	ethical conduct	CT less important (1 -2)	27%	22%	29%	20%	30%	22%
		equal importance (3)	46%	58%	48%	52%	45%	49%
		CT more important (4-5)	26%	19%	23%	28%	25%	29%

Table C.3 continued

Item	Item description	Options	Discipline: Science (n=110)	Discipline: History (n=36)	Education setting: High school (n=96)	Education setting: Tertiary (n=50)	Science: Education setting: High school (n=69)	Science: Education setting: Tertiary (n=41)
			Proportions by demographic factors#					
D_2f	ICT	CT less important (1 -2)	13%	6%	13%	8%	14%	10%
		equal importance (3)	41%	39%	44%	34%	45%	34%
		CT more important (4-5)	46%	56%	44%	58%	41%	56%
D_2g	intercultural understanding	CT less important (1 -2)	8%	17%	9%	12%	6%	12%
		equal importance (3)	35%	61%	47%	32%	41%	27%
		CT more important (4-5)	56%	22%	44%	56%	54%	61%
D_2h	knowledge	CT less important (1 -2)	24%	28%	29%	16%	29%	15%
		equal importance (3)	50%	44%	45%	56%	45%	59%
		CT more important (4-5)	26%	28%	26%	28%	26%	27%
D_2i	literacy	CT less important (1 -2)	40%	42%	39%	44%	41%	39%
		equal importance (3)	50%	50%	55%	40%	54%	44%
		CT more important (4-5)	10%	8%	6%	16%	6%	17%
D_2j	numeracy	CT less important (1 -2)	32%	14%	31%	20%	36%	24%
		equal importance (3)	55%	39%	51%	50%	55%	54%
		CT more important (4-5)	14%	47%	18%	30%	9%	22%

Table C.3 continued

Item	Item description	Options	Discipline: Science (n=110)	Discipline: History (n=36)	Education setting: High school (n=96)	Education setting: Tertiary (n=50)	Science: Education setting: High school (n=69)	Science: Education setting: Tertiary (n=41)
			Proportions by demographic factors#					
D_2k	planning	CT less important (1 -2)	16%	6%	15%	12%	19%	12%
		equal importance (3)	53%	53%	57%	44%	55%	49%
		CT more important (4-5)	31%	42%	28%	44%	26%	39%
D_2l	problem solving	CT less important (1 -2)	26%	22%	26%	24%	28%	24%
		equal importance (3)	70%	75%	73%	68%	71%	68%
		CT more important (4-5)	4%	3%	1%	8%	1%	7%
D_2m	reasoning	CT less important (1 -2)	24%	25%	26%	20%	25%	22%
		equal importance (3)	75%	69%	72%	78%	75%	76%
		CT more important (4-5)	1%	6%	2%	2%	0%	2%
D_2n	teamwork	CT less important (1 -2)	15%	8%	19%	2%	22%	2%
		equal importance (3)	42%	28%	42%	32%	45%	37%
		CT more important (4-5)	44%	64%	40%	66%	33%	61%
D_2o	understanding	CT less important (1 -2)	29%	25%	27%	30%	28%	32%
		equal importance (3)	55%	67%	58%	56%	57%	51%
		CT more important (4-5)	16%	8%	15%	14%	16%	17%

due to rounding to whole numbers proportions may be ± 1% of the sum of the displayed values.

Table C.4 Educator responses and corresponding coding themes for Question 3c: Please explain how you know if critical thinking skills are an emergent property of your students' learning?

demography			
discipline	**education setting**	**response**	**themes**
science	hs	Through discussions, assessment tasks and practical activities where conversation is involved.	evident in interactions with student through general class activities and assessment
science	hs	They are asking more questions about what they are being taught, in terms of what if…. isn't this….. why then would…..	evident in interactions with students
science	hs	I dont formally know since we dont directly assess it. I do see glimpses/elements of critical thinking is some of the work that students do in class and on assessment	evident in interactions with students through general class activities and assessment
science	hs	Incidental to problem solving and in proposing predictions. Proposing solutions to issues	through general class activities and assessment
science	ter	You set them off to evaluate positions and claims with diverging viewpoints, and see how they deal with disparate opinions expressed by others. You can set them a bunch of papers and ask them to evaluate how good the evidence is in each.	evident in interactions with students through evaluation tasks
science	ter	Students are taught to learn and research for themselves, and encouraged not to believe everything they read without the proper supporting references/evidence.	through general class activities and assessment infer the use of evaluation style-tasks
science	ter	Students are able to solve problems not part of the curriculum.	evident in interactions with students

Table C.5 Proportions of educators who indicated per Likert-option groupings for Question 4. Displayed by demographic factors: discipline, education setting, and by education setting within the science cohort.

			Proportions by demographic factors					
			Discipline		Education setting		Science: Education setting	
Item	Item description	Options	Science (n=110)	History (n=36)	High school (n=96)	Tertiary (n=50)	High school (n=69)	Tertiary (n=41)
D_4a	problem based learning effectiveness	ineffective (1 -2)	3%	0%	2%	2%	3%	2%
		neither effective or ineffective (3)	7%	19%	11%	8%	7%	7%
		effective (4-5)	84%	72%	84%	74%	88%	76%
		no experience	6%	8%	2%	16%	1%	15%
D_4b	collaborative learning effectiveness	ineffective (1 -2)	3%	17%	16%	14%	14%	15%
		neither effective or ineffective (3)	7%	22%	26%	38%	28%	41%
		effective (4-5)	84%	58%	56%	34%	55%	29%
		no experience	6%	3%	2%	14%	3%	15%
D_4c	concept mapping effectiveness	ineffective (1 -2)	15%	22%	23%	4%	22%	2%
		neither effective or ineffective (3)	21%	28%	27%	14%	25%	15%
		effective (4-5)	39%	36%	43%	30%	48%	24%
		no experience	25%	14%	7%	52%	6%	59%
D_4d	logic models effectiveness	ineffective (1 -2)	5%	17%	11%	2%	9%	0%
		neither effective or ineffective (3)	19%	14%	20%	14%	20%	17%
		effective (4-5)	34%	11%	34%	16%	42%	20%
		no experience	42%	58%	34%	68%	29%	63%
D_4e	context dependant sets effectiveness	ineffective (1 -2)	13%	11%	13%	12%	13%	12%
		neither effective or ineffective (3)	21%	19%	26%	10%	28%	10%
		effective (4-5)	21%	19%	23%	16%	22%	20%
		no experience	45%	50%	39%	62%	38%	59%

Table C.5 continued

Item	Item description	Options	Discipline: Science (n=110)	Discipline: History (n=36)	Education setting: High school (n=96)	Education setting: Tertiary (n=50)	Science: Education setting: High school (n=69)	Science: Education setting: Tertiary (n=41)
D_4f	debating effectiveness	ineffective (1 -2)	12%	6%	11%	8%	13%	10%
		neither effective or ineffective (3)	15%	3%	14%	10%	19%	10%
		effective (4-5)	47%	92%	64%	48%	52%	39%
		no experience	25%	0%	11%	34%	16%	41%
D_4g	flipped classroom effectiveness	ineffective (1 -2)	17%	14%	18%	14%	20%	12%
		neither effective or ineffective (3)	16%	17%	18%	14%	16%	17%
		effective (4-5)	19%	22%	21%	18%	22%	15%
		no experience	47%	47%	44%	54%	42%	56%
D_4h	scientific method effectiveness	ineffective (1 -2)	5%	17%	6%	12%	4%	7%
		neither effective or ineffective (3)	9%	17%	10%	12%	6%	15%
		effective (4-5)	80%	11%	68%	54%	88%	66%
		no experience	5%	56%	16%	22%	1%	12%
D_4i	Socratic questioning effectiveness	ineffective (1 -2)	5%	6%	8%	0%	9%	0%
		neither effective or ineffective (3)	17%	11%	17%	14%	17%	17%
		effective (4-5)	33%	58%	40%	38%	33%	32%
		no experience	45%	25%	35%	48%	41%	51%
D_4j	peer assessment effectiveness	ineffective (1 -2)	30%	31%	35%	20%	33%	24%
		neither effective or ineffective (3)	25%	11%	21%	22%	25%	24%
		effective (4-5)	27%	44%	34%	26%	32%	20%
		no experience	18%	14%	9%	32%	10%	32%

Table C.5 continued

Item	Item description	Options	Discipline: Science (n=110)	Discipline: History (n=36)	Education setting: High school (n=96)	Education setting: Tertiary (n=50)	Science: Education setting: High school (n=69)	Science: Education setting: Tertiary (n=41)
D_4k	inductive reasoning activities	ineffective (1 -2)	5%	3%	6%	0%	7%	0%
		neither effective or ineffective (3)	17%	14%	21%	8%	23%	7%
		effective (4-5)	45%	39%	45%	42%	46%	44%
		no experience	33%	44%	28%	50%	23%	49%
D_4l	deductive reasoning effectiveness	ineffective (1 -2)	5%	6%	6%	4%	7%	2%
		neither effective or ineffective (3)	18%	8%	18%	12%	22%	12%
		effective (4-5)	55%	53%	60%	42%	59%	46%
		no experience	22%	33%	16%	42%	12%	39%
D_4m	constucting critiques effectiveness	ineffective (1 -2)	8%	6%	9%	4%	10%	5%
		neither effective or ineffective (3)	15%	25%	18%	16%	13%	17%
		effective (4-5)	42%	58%	48%	42%	46%	34%
		no experience	35%	11%	25%	38%	30%	44%
D_4n	class discussion effectiveness	ineffective (1 -2)	18%	3%	16%	12%	20%	15%
		neither effective or ineffective (3)	17%	19%	19%	16%	17%	17%
		effective (4-5)	62%	78%	66%	66%	62%	61%
		no experience	3%	0%	0%	6%	0%	7%
D_4o	oral presentation effectiveness	ineffective (1 -2)	31%	25%	32%	24%	35%	24%
		neither effective or ineffective (3)	44%	39%	48%	32%	48%	37%
		effective (4-5)	15%	31%	16%	26%	13%	20%
		no experience	10%	6%	4%	18%	4%	20%

Proportions by demographic factors

Table C.5 continued

Item	Item description	Options	Proportions by demographic factors: Discipline: Science (n=110)	History (n=36)	Education setting: High school (n=96)	Tertiary (n=50)	Science: Education setting: High school (n=69)	Tertiary (n=41)
D_4p	case studies effectiveness	ineffective (1 -2)	16%	6%	16%	10%	19%	12%
		neither effective or ineffective (3)	20%	25%	27%	10%	25%	12%
		effective (4-5)	50%	61%	50%	58%	48%	54%
		no experience	14%	8%	7%	22%	9%	22%
D_4q	writing tasks effectiveness	ineffective (1 -2)	27%	3%	29%	6%	39%	7%
		neither effective or ineffective (3)	29%	36%	30%	32%	28%	32%
		effective (4-5)	33%	53%	34%	44%	28%	41%
		no experience	11%	8%	6%	18%	6%	20%
D_4r	argument analysis effectiveness	ineffective (1 -2)	7%	3%	8%	2%	10%	2%
		neither effective or ineffective (3)	8%	8%	10%	4%	10%	5%
		effective (4-5)	55%	83%	65%	56%	58%	49%
		no experience	30%	6%	17%	38%	22%	44%

Table C.6 Proportions of educators who chose each option for Question 5 items. Displayed by demographic factors: discipline, education setting, and by education setting within the science cohort.

Assessment approach		Demographic factor					
		Education setting		Science: Education setting		Discipline	
Item theme and code	Options	Tertiary (n=50)	High school (n=96)	Tertiary (n=41)	High school (n=69)	Science (n=110)	History (n=36)
Course evaluation form (A_5a)	Self - developed item	34%	59%	34%	58%	49%	56%
	External resource	18%	25%	15%	28%	23%	22%
	Would not use	40%	21%	41%	22%	29%	22%
	Skipped	10%	3%	12%	1%	5%	6%
Discipline specific CT skills test (A_5b)	Self - developed item	60%	40%	61%	36%	45%	50%
	External resource	30%	56%	37%	59%	51%	36%
	Would not use	18%	14%	12%	13%	13%	22%
	Skipped	2%	2%	2%	1%	2%	3%
Extended response items (A_5c)	Self - developed item	56%	70%	49%	65%	59%	83%
	External resource	10%	47%	10%	54%	37%	25%
	Would not use	24%	1%	27%	1%	11%	3%
	Skipped	12%	4%	15%	4%	8%	3%
General CT skills test (A_5d)	Self - developed item	26%	25%	29%	23%	25%	25%
	External resource	32%	56%	34%	61%	51%	39%
	Would not use	40%	21%	34%	20%	25%	33%
	Skipped	8%	6%	10%	6%	7%	6%
Multiple ranking items (A_5e)	Self - developed item	18%	20%	20%	20%	20%	17%
	External resource	14%	41%	17%	45%	35%	22%
	Would not use	52%	35%	44%	33%	37%	53%
	Skipped	18%	10%	22%	10%	15%	8%
Multiple choice items (A_5f)	Self - developed item	34%	30%	37%	30%	33%	28%
	External resource	16%	45%	20%	51%	39%	22%
	Would not use	42%	30%	34%	28%	30%	47%
	Skipped	10%	5%	12%	4%	7%	6%
Rubric (A_5g)	Self - developed item	50%	70%	51%	68%	62%	67%
	External resource	14%	32%	12%	36%	27%	22%
	Would not use	28%	13%	24%	12%	16%	22%
	Skipped	12%	2%	15%	1%	6%	3%
Self - report items (A_5h)	Self - developed item	38%	50%	34%	49%	44%	53%
	External resource	10%	24%	10%	28%	21%	14%
	Would not use	42%	26%	41%	25%	31%	33%
	Skipped	12%	6%	15%	7%	10%	3%

Table C.7 Summary of group dissimilarity analysis from the SIMPER approach for demographic factors identified as significant in the assessment sub-profile. Displayed in rank order of contributions to the differences by education setting (i.e. between high school and tertiary educators).

		SIMPER rank (% contribution of item to difference between groups; group with higher mean score)			
Item theme	**Item code**	**Education setting (HS vs TER)**		**Science: Education setting (Sc HS v Sc TER)**	
self-developed course evaluation form	**A_5a - SD**	**1**	**(4.59%; HS)**	4	(4.49%; Sc HS)
self-developed discipline specific CT test	A_5b - SD	2	(4.57%; TER)	2	(4.58%; Sc TER)
externally sourced discipline specific CT test	A_5b - ER	3	(4.54%; HS)	3	(4.52%; Sc HS)
externally sourced general CT test	**A_5d - ER**	4	(4.54%; HS)	**1**	**(4.59%; Sc HS)**
self developed rubric	A_5g - SD	5	(4.43%; HS)	7	(4.32%; Sc HS)
would not use multiple ranking items	A_5e - NO	6	(4.36%; TER)	10	(4.06%; Sc TER)
self-developed self-report items	A_5h - SD	7	(4.33%; HS)	9	(4.23%; Sc HS)
self-developed extended response items	A_5c - SD	8	(4.21%; HS)	6	(4.34%; Sc HS)
would not use multiple choice items	A_5f - NO	9	(4.03%; TER)	15	(3.64%; Sc TER)
externally sourced extended response items	A_5c - ER	10	(3.98%; HS)	5	(4.42%; Sc HS)
would not use self-report items	A_5h - NO	11	(3.98%; TER)	12	(3.85%; Sc TER)
externally sourced multiple choice item	A_5f - ER	12	(3.93%; HS)	8	(4.23%; Sc HS)
would not use course evalaution form	A_5a - NO	13	(3.77%; TER)	14	(3.78%; Sc TER)
self-developed multiple choice items	A_5f - SD	14	(3.77%; TER)	13	(3.79%; Sc TER)
would not use general CT test	A_5d - NO	15	(3.76%; TER)	17	(3.41%; Sc TER)
externally sourced multiple-ranking items	A_5e - ER	16	(3.69%; HS)	11	(3.93%; Sc HS)
I look for specific evidence of CT embedded in students work.	A_6d	17	(3.53%; TER)	16	(3.56%; Sc TER)
self-developed general CT skills tests	A_5d - SD	18	(3.32%; TER)	18	(3.34%; Sc TER)
would not use rubric	A_5g - NO	19	(3.14%; TER)	20	(2.87%; Sc TER)
externally sourced rubric	A_5g - ER	20	(3.06%; HS)	19	(3.23%; Sc HS)
externally sourced course evaluation form	A_5a - ER	21	(2.87%; HS)	21	(2.85%; Sc HS)
self-developed multiple ranking items	A_5e - SD	22	(2.55%; HS)	23	(2.57%; Sc HS)
externally sourced self-report items	A_5h - ER	23	(2.43%; HS)	22	(2.63%; Sc HS)
would not use discipline specific CT test	A_5b - NO	24	(2.27%; TER)	25	(1.87%; Sc HS)
would not use extended response items	A_5c - NO	25	(2.2%; TER)	24	(2.41%; Sc TER)
I use rubrics to assess CT in written responses.	A_6g	26	(1.63%; HS)	27	(1.42%; Sc HS)
I do not look for specific evidence of CT when assessing students work, but know my students are developing this skill as their	A_6c	27	(1.55%; TER)	26	(1.51%; Sc TER)
I use a combination of rubrics, self-report and multiple-choice items to assess CT.	A_6h	28	(1.23%; HS)	28	(1.28%; Sc HS)
I do not look for specific evidence of CT when assessing students work.	A_6b	29	(1.02%; TER)	29	(1.21%; Sc TER)
I assess CT development using self report assessment items.	A_6e	30	(0.91%; HS)	30	(0.99%; Sc HS)
I have difficulty identifying student work that reflects CT.	A_6a	31	(0.75%; HS)	31	(0.91%; Sc HS)
I assess CT development using multiple choice items assessment items.	A_6f	32	(0.55%; HS)	32	(0.6%; Sc HS)
I assess CT development using published CT test items.	A_6i	33	(0.49%; HS)	33	(0.58%; Sc HS)

Table C.8 Summary of items that contributed at least 2% in the group dissimilarity analysis from the SIMPER approach for demographic factors identified as significant in the global profile. Items displayed in alphabetical order. Item in bold indicates the highest contributing item to group differences for that interaction pair.

Item code	Item theme	State/education setting interaction effect pairing			
		VIC: high school vs tertiary	NSW: high school vs tertiary	High school: SA vs VIC	High school: SA v WA
D_3a	approaches involving explicit instruction			x	x
D_3b	approaches embedded across the curriculum		x	x	x
D_4f	debating	x			x
D_4i	Socratic questioning			x	x
D_4k	inductive reasoning activities	x			
D_4l	deductive reasoning	x			
D_4r	**argument analysis**	**x**			
A_5a - ER	use external resource: course evalution form			x	x
A_5a - NO	would not use a course evaluation form		x		
A_5a - SD	**self-developed course evaluation form**	x	**x**	x	x
A_5b - ER	use external resource: discipline specific CT skills test	x	x	x	x
A_5b - SD	self-developed discipline specific CT skills test	x	x		x
A_5c - ER	use external resource: extended response items	x		x	x
A_5c - NO	would not use extended response items	x			
A_5c - SD	self-developed extended response items	x	x	x	x
A_5d - ER	use external resource: general CT skills test	x	x	x	x
A_5d - NO	would not use general CT skills test				x
A_5e - ER	use external resource: multiple ranking items		x	x	x
A_5e - NO	would not use multiple ranking items	x	x	x	x
A_5e - SD	self-developed multiple ranking items				x
A_5f - ER	use external resource: multiple choice items	x	x	x	x
A_5f - NO	would not use multiple choice items	x	x	x	
A_5f - SD	self-developed multiple choice items				x
A_5g - ER	use external resource: rubric			x	x
A_5g - SD	**self-developed rubric**	x	x	**x**	**x**
A_5h - NO	would not use self-report items	x		x	
A_5h - SD	self-developed self-report items	x	x	x	x
A_6d	ok for specific evidence of CT embedded in students work.	x		x	

APPENDIX D Case study perception survey template and statement key

Q1). Please indicate how often you think [insert course code] gave you the opportunity to practice the following... (used in the post-test only)

Statement options: (Shade the relevant box for each statement)

Every class	Weekly	A few times	Once	Never
☐	☐	☐	☐	☐

Table D.1 Student perception of learning opportunities (practice) statement key. Statements only included in the post-test. Statements shaded in Blue relate to CAT-measurable skills.

#	Statement
Q1	summarize a pattern of information without making inappropriate inferences.
Q2	evaluate how strongly correlational-type data supports a hypothesis.
Q3	provide alternative explanations for observations.
Q4	identify additional information needed to evaluate a hypothesis or particular explanation of an observation.
Q5	evaluate whether spurious relationships strongly support a claim.
Q6	provide alternative explanations for spurious relationships.
Q7	identify additional information needed to evaluate a hypothesis or particular explanation of an observation.
Q8	determine whether an inference in an advertisement is supported by information.
Q9	provide relevant alternative interpretations of information.
Q10	separate relevant from irrelevant information when solving a real-world problem.
Q11	analyse and integrate information from separate sources to solve a real-world problem.
Q12	use basic mathematical skills to help solve a real-world problem.
Q13	identify suitable solutions for a real-world problem using relevant information.
Q14	identify and explain the best solution for a real-world problem using relevant information.
Q15	explain how changes in a real-world problem situation might affect the solution.
Q16	explain the methods of science.
Q17	consider why scientific knowledge is testable.
Q18	explain why scientific knowledge is testable by further inquiry.
Q19	explain the role of science in society.
Q20	explain the relevance of science in society.
Q21	design and plan an investigation
Q22	collect and accurately record scientific data.
Q23	interpret and draw conclusions from scientific data.

Q2). Please indicate your level of agreement with the following statements to tell us how well you think you can do each skill. (Note: B&S1 had no pre-test perception survey and the statement they were asked told them to indicate which skills they thought the course had helped them develop)

Statement options: (Shade the relevant box for each statement)

Strongly agree	Agree	Neutral	Disagree	Strongly disagree
☐	☐	☐	☐	☐

Table D.2 Student skill statement key. Statements shaded in Blue relate to CAT-measurable skills.

#	Statement
Q1	I can summarize a pattern of information without making inappropriate inferences.
Q2	I can evaluate how strongly correlational-type data supports a hypothesis.
Q3	I can provide alternative explanations for observations.
Q4	I can identify additional information needed to evaluate a hypothesis or particular explanation of an observation.
Q5	I can evaluate whether spurious relationships strongly support a claim.
Q6	I can provide alternative explanations for spurious relationships.
Q7	I can identify additional information needed to evaluate a hypothesis or particular explanation of an observation.
Q8	I can determine whether an inference in an advertisement is supported by information.
Q9	I can provide relevant alternative interpretations of information.
Q10	I can separate relevant from irrelevant information when solving a real-world problem.
Q11	I can analyse and integrate information from separate sources to solve a real-world problem.
Q12	I can use basic mathematical skills to help solve a real-world problem.
Q13	I can identify suitable solutions for a real-world problem using relevant information.
Q14	I can identify and explain the best solution for a real-world problem using relevant information.
Q15	I can explain how changes in a real-world problem situation might affect the solution.
Q16	I can explain the methods of science.
Q17	I can explain why current scientific knowledge is contestable.
Q18	I can explain why scientific knowledge is testable by further inquiry.
Q19	I can explain the role of science in society.
Q20	I can explain the relevance of science in society.
Q21	I can design and plan an investigation
Q22	I can collect and accurately record scientific data.
Q23	I can interpret and draw conclusions from scientific data.

Q3). Please respond to the following statements concerning your attributes and dispositions

(Note: B&SV1 only had a post-test).

Statement options: (Shade the relevant box for each statement)

Strongly agree	Agree	Neutral	Disagree	Strongly disagree
☐	☐	☐	☐	☐

Table D.3 Student attribute and disposition statement key. Statements Q24-Q27 (shaded in grey) were included in all student surveys, whereas statements Q28-Q38 were only included in B&SV2.

#	Statement
Q24	I am in the habit of connecting key ideas I learn in my classes with other knowledge.
Q25	I am in the habit of applying what I learn in classes to other situations.
Q26	I am in the habit of using systematic reasoning in my approach to problems.
Q27	I use a critical approach to analyzing data and arguments in my daily life.
Q28	I usually try to think about the bigger picture during a discussion.
Q29	I often use new ideas to shape (modify) the way I do things.
Q30	I often re-evaluate my experiences so that I can learn from them.
Q31	I use more than one source to find out information for myself.
Q32	I am often on the lookout for new ideas.
Q33	I usually check the credibility of the source of information before making judgements.
Q34	I sometimes find a good argument that challenges some of my firmly held beliefs.
Q35	It is important to understand other people's viewpoint on an issue.
Q36	I usually think about the wider implications of a decision before taking action.
Q37	It is important to justify the choices I make.
Q38	I often think about my actions to see whether I could improve them.

Demographic information requested

Gender: Male Female Other Prefer not to disclose

Degree: open-ended

Student ID: (for generating pre-test and post-test matches)

APPENDIX E Case study observation tools

Classroom Observation Protocol for Undergraduate STEM (COPUS) (Smith et al., 2013).

Table E.1 COPUS codes and descriptions. *Note*. Reprinted from Classroom Observation Protocol for Undergraduate STEM (Smith et al., 2013).

Removed due to copyright restriction

Table E.2 An excerpt of the COPUS coding form. Observers check off each code that occurs during each 2-minute block. *Note*. Reprinted from Classroom Observation Protocol for Undergraduate STEM (Smith et al., 2013).

Removed due to copyright restriction

Electronic Quality of Inquiry Protocol (EQUIP) (Marshall et al., 2009).

Table E.3 Excerpt of EQUIP protocol aspects used in both case studies. ***Note.*** **Adapted from Marshall et al. (2009).**

EQUIP component		Level	Description
Cognitive Levels			Removed due to copyright restriction
Discourse factors	Questioning Level		
	Complexity of Questions		
	Classroom Interactions		

Table E.4 Excerpt of additional EQUIP protocol aspects that were used in the biology and society case study method. ***Note.*** **Adapted from Marshall et al. (2009).**

EQUIP component		Level	Description
Instructional factors	Teacher role		Removed due to copyright restriction
	Student role		

Queensland School Reform Longitudinal Study protocol (QSRLS) (Lingard et al., 2001; The School Reform Longitudinal Study Research Team, 2001).

Table E.5 Scoring instructions for the QSRLS protocol dimension: Knowledge integration. *Note*: Excerpt from The School Reform Longitudinal Study Research Team (The School Reform Longitudinal Study Research Team, 2001, p.3)

Removed due to copyright restriction

Table E.6 Scoring instructions for the QSRLS protocol dimension: Problematic Knowledge: Construction of knowledge. *Note*: Excerpt from The School Reform Longitudinal Study Research Team (The School Reform Longitudinal Study Research Team, 2001, p.4).

Removed due to copyright restriction

Table E.7 Scoring instructions for the QSRLS protocol dimension: Problem-based curriculum. ***Note*****: Excerpt from The School Reform Longitudinal Study Research Team (The School Reform Longitudinal Study Research Team, 2001, p.19).**

Removed due to copyright restriction

APPENDIX F CAT instrument information

Sample CAT question. ***Note*****: Reprinted from CAT App PowerPoint (Centre for Assessment and Teaching, 2019a, p.11).**

Removed due to copyright restriction

Example of scoring approach for sample questions

Question 1.

Question 2.

Content removed for privacy reasons

Questions 3.

Table F.1 List of CAT questions and associated skills. ***Note*.** **Adapted from CAT Training Manual (Centre for Assessment and Learning, 2013, p.23).**

Content removed for privacy reasons

APPENDIX G Chemistry course learning outcomes and graduate attributes

Table G.1 Alignment of assessment items to expected course learning outcomes, as per the chemistry 2015 Statement of Assessment Methods. ***Note.*** **Reprinted from 'Statement of Assessment Methods -2015' (Koeper, 2015, p.2)**

Removed due to copyright restriction

Table G.2 Alignment of assessment items to associated graduate qualities, as per the chemistry 2015 Statement of Assessment Methods. *Note.* Reprinted from 'Statement of Assessment Methods -2015' (Koeper, 2015, p.2-3).

Removed due to copyright restriction

APPENDIX H Additional results from Chapter 4: Exploring CT Development in Chemistry

Table H.1 Summary of the CAT results for the chemistry course (n=69). Due to variations in the scales across questions, different statistical tests were applied as appropriate. These are indicated through the symbols near the question number and are explained below in the table notes.

	Q1†		Q2^		Q3^		Q4^	
	pre	post	pre	post	pre	post	pre	post
Maximum	1	1	3	3	3	3	4	4
Median	0	1	1	2	1	1	1	1
Mean	0.46	0.51	1.38	1.54	1.19	1.30	0.86	1.07
Sum	32	35	95	106	82	90	59	74
difference	3		11		8		15	
Statistical test statistic	0.261		122		132		212	
Statistical test: p-value§	0.609		0.34		0.293		0.107	

	Q5†		Q6^		Q7^		Q8†	
	pre	post	pre	post	pre	post	pre	post
Maximum	1	1	3	3	2	2	1	1
Median	1	1	2	1	0	0	1	1
Mean	0.51	0.68	1.52	1.48	0.46	0.41	0.64	0.70
Sum	35	47	105	102	32	28	44	48
difference	**12**		-3		-4		4	
Statistical test statistic	**4.328**		-50		-74		0.521	
Statistical test: p-value§	**0.037**		0.672		0.636		0.47	

	Q9^		Q10^		Q11^		Q12†	
	pre	post	pre	post	pre	post	pre	post
Maximum	2	2	4	4	2	2	1	1
Median	1	1	3	3	1	1	1	1
Mean	0.74	0.88	2.94	3.23	0.93	1.22	0.75	0.78
Sum	51	61	203	223	64	84	52	54
difference	10		**20**		**20**		2	
Statistical test statistic	148		**336**		**316**		0.162	
Statistical test: p-value§	0.153		**0.019**		**0.003**		0.686	

	Q13^		Q14^		Q15^		total#	
	pre	post	pre	post	pre	post	pre	post
Maximum	3	3	5	5	3	3	32	31
Median	1	1	3	3	0	1	16	18
Mean	1.17	1.25	2.19	2.43	0.75	0.84	16.49	18.32
Sum	81	86	151	168	52	58	1138	1264
difference	5		17		6		**126**	
Statistical test statistic	102		119		76		**3.469**	
Statistical test: p-value§	0.502		0.374		0.444		**0.0009**	

† chi-squared test statistic (chi-squared), with two-sided p-value. A contingency test comparing the proportions of 0 and 1 scores between pre and post.

^ Wilcoxon signed rank test statistic (*W*), with two-sided p-value. A non-parametric version of the paired t-test, to assess whether matched population mean ranks differ between pre and post.

paired t-test statistic (*t*), with two-sided p-value. A parametric test after confirmation of normality (using the Shapiro-Wilk and D'Agostino & Pearson normality tests) to assess whether population means differ pre and post.

§ In all cases, alpha was set to 0.05, where bolded p-values indicate where a significant difference lies.

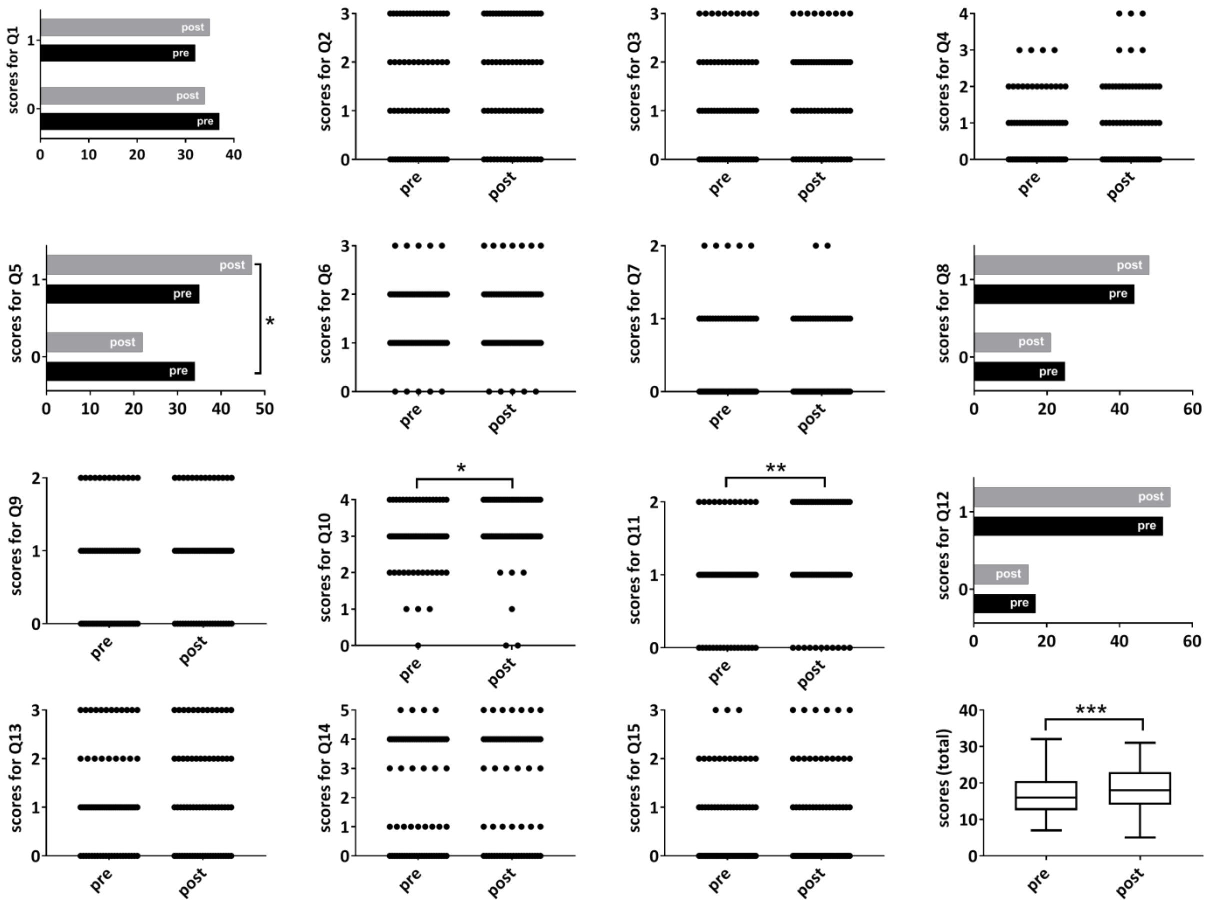


Figure H.1 Summary of the chemistry cohort's pre-test and post-test CAT results (n=69). In all cases, alpha was set to 0.05, where an asterisk indicates a result where a significant difference lies and the number of asterisks denotes the level of significance, * <0.05, ** <0.01, * <0.001.**

Table H.2 Chemistry cohort CAT pre-test results compared to the national results for US freshman. ***Note.*** **Reprinted from Flinders University CAT Institution report (Centre for Assessment and Improvement of Learning, 2019b, p.24).**

Removed due to copyright restriction

Table H.3 Chemistry cohort CAT post-test results compared to the national results for US freshman. ***Note.*** **Reprinted from Flinders University CAT Institution report (Centre for Assessment and Improvement of Learning, 2019b, p.39).**

Removed due to copyright restriction

Table H.4 Summary of changes in student perceptions and performance, in the context of student perceptions about the CT learning opportunities in chemistry. Students grouped by the extent to which they perceived they had the opportunity to practice each skill, where the % value indicates the proportion of students who perceived that extent of practice. The symbols (↑ = ↓) displayed in the change in perception and performance columns represent the mode change across the semester, where "↑" represents an increase in their perception and/or performance, "=" no change in their perception and/or performance, and "↓" indicates a decrease/reduction in that group's perception and/or performance. Bolded question and skills reflect statistically significant differences in performance on that question/skill in the CAT test.

			Practised regularly			Practised irregularly			Never practised		
Question	CT skill description (based on CAT measurable skills)	n	%	Change in perception of ability	Change in CAT performance	%	Change in perception of ability	Change in CAT performance	%	Change in perception of ability	Change in CAT performance
1	summarize a pattern of information without making inappropriate inferences.	62	59.7	=	=	40.3	=	=	0.0	n/a	n/a
2	evaluate how strongly correlational-type data supports a hypothesis.	62	47.6	=	=	47.6	=	=	4.8	↓	=
3	provide alternative explanations for observations.	62	50.8	=	~	49.2	=	=	0.0	n/a	n/a
4	identify additional information needed to evaluate a hypothesis or particular explanation of an observation.	62	55.6	=	=	44.4	=	=	0.0	n/a	n/a
5	**evaluate whether spurious relationships strongly support a claim.***	**60**	**33.9**	↑	=	**59.7**	↑	=	**6.5**	↓	=
6	provide alternative explanations for spurious relationships.	62	25.4	=	=	66.7	=	=	7.9	=	↑
7	identify additional information needed to evaluate a hypothesis or particular explanation of an observation.	62	55.6	=	=	44.4	=	=	0.0	n/a	n/a
8	determine whether an inference in an advertisement is supported by information.	61	35.5	=	=	43.5	^	=	21.0	=	=
9	provide relevant alternative interpretations of information.	61	53.2	=	=	45.2	=	=	1.6	↓	=
10	**separate relevant from irrelevant information when solving a real-world problem.**	**62**	**66.7**	=	↑	**31.7**	=	~	**1.6**	↑	=
11	**analyse and integrate information from separate sources to solve a real-world problem.***	**60**	**58.1**	=	=	**38.7**	^	=	**3.2**	~	~
12	use basic mathematical skills to help solve a real-world problem.	63	88.9	=	=	11.1	=	=	0.0	n/a	n/a
13	identify suitable solutions for a real-world problem using relevant information.	62	65.1	=	=	33.3	=	=	1.6	↓	↑
14	identify and explain the best solution for a real-world problem using relevant information.	61	59.7	=	=	37.1	=	=	3.2	^	~
15	explain how changes in a real-world problem situation might affect the solution.	62	58.7	=	=	38.1	=	=	3.2	↓	^

~BIMODAL- equal number of no change and increase in performance; *Note for these item N=60 which means 8 students who were included in the pretest posttest CAT analysis failed to supply a response to this question on either the perception survey pretest or posttest; ^BIMODAL- equal number of no change and decrease in performance.

APPENDIX I Biology and society course learning outcomes and graduate attributes

Table I.1 Alignment of assessment items to expected course learning outcomes as per the biology and society 2015 Statement of Assessment Methods. ***Note.*** **Reprinted from 'Statement of Assessment Methods -2015' (Hunter, 2015, p.2).**

Removed due to copyright restriction

Table I.2 Alignment of assessment items to associated graduate qualities as per the biology and society 2015 Statement of Assessment Methods. ***Note.*** **Reprinted from 'Statement of Assessment Methods -2015' (Hunter, 2015, p.2-3).**

Removed due to copyright restriction

APPENDIX J Biology and society version 2 intervention tasks

Table J.1 List of skill sets associated with each CAT question. *Note*. Adapted from Center for Assessment and Improvement of Learning (2019a, p.25-26).

	CT skills associated with CAT Questions
Skill set 1	Q1: Summarize a pattern of information without making inappropriate inferences.
	Q2: Evaluate how strongly correlational-type data supports a hypothesis.
	Q3: Provide alternative explanations for observations.
	Q4: Identify additional information needed to evaluate a hypothesis or particular explanation of an observation.
	Q5: Evaluate whether spurious relationships strongly support a claim.
	Q6: Provide alternative explanations for spurious relationships.
	Q7: Identify additional information needed to evaluate a hypothesis/interpretation.
	Q8: Determine whether an invited inference in an advertisement is supported by information.
	Q9: Provide relevant alternative interpretations of information.
Skill set 2	Q10: Separate relevant from irrelevant information when solving a real-world problem.
	Q11: Analyze and integrate information from separate sources to solve a real-world problem.
	Q12: Use basic mathematical skills to help solve a real-world problem.
	Q13: Identify suitable solutions for a real-world problem using relevant information.
	Q14: Identify and explain the best solution for a real-world problem using relevant information.
	Q15: Explain how changes in a real-world problem situation might affect the solution.

Intervention task 1 - comprising two tasks set as an individual thinking question, with part a completed at the start of class (followed by a group discussion) and part b completed at the end of class.

The first intervention focussed on getting students to practice "evaluating whether spurious relationships strongly support a claim". Two scenarios were developed, one concentrating on food allergies and the case for food labelling; the other focusing on GMO crops and the potential for crop contamination. The scenarios were developed using CAT methodology appropriate to a skill set one development, an approach that essentially involves creating an opportunity for students to practice identifying the strength of an argument and generating alternative interpretations and conclusions about the information. The first scenario was given at the start of class, followed by an instructor-led discussion about the elements of an argument. The session also included content about what constitutes valid and reliable evidence, as well as an explanation of the difference between correlation and causation. The focus of class activities then shifted into applying this line of thinking to popular GMO issues, with the last component of the session involving students individually completing the second intervention question.

a) Scenario: A nutritionist reported that food allergies are increasing at an alarming rate, and attributes the cause to the addition of GMO's to our diets. The nutritionist believes that labelling products containing GMO should be compulsory, as it will help consumers to identify ingredients that cause allergies. The nutritionist proposed the following reasons for the compulsory labelling GMO foods: -

- There has been a 50% increase in food allergies in children between 1997-2011.
- GMO's have become an increasingly available ingredient over the same time period, with the number of hectares of GMO crops planted increasing from 1.7million (1996) to 160 million (2011).
- A food allergy sends someone to the ER every 3 minutes.

Do the reasons proposed by the nutritionist provide strong support for the compulsory labelling of GMO food? Explain your answer.

b) Scenario: *An organic farmer is facing criminal charges for stealing his neighbours genetically modified pest-resistant corn crop. During the legal proceedings, the judge is presented with the following evidence: -*

- *Genetic testing of the organic farmer's corn crop revealed that some of the corn contained the pest-resistant gene.*
- *The organic farmer has been seen walking around his property late at night, and one witness reported that they saw the organic farmer go into his neighbour's cornfield.*
- *The pest-resistant crop and the organic crop are planted in fields next to each other.*

Does the evidence provided clearly support a guilty verdict? Explain your answer.

Intervention task 2 - part a set as an individual thinking question; and part b - d set as follow up group discussion activity

The second CT intervention was given to the students in week 8. Students were asked to complete these questions during lecture time,[29] and the individual component of the intervention replaced their regular participation quiz. The intervention questions were written in the style of a skill set 2 CAT questions, targeting skills involved in real-world problem solving including separating relevant and irrelevant information, considering alternatives and explaining the best solution for a problem. To practice these skills, students were given a scenario containing a mix of

[29] Students who were not present at the lecture were asked to complete the individual task in place of a lecture summary.

important and distracting information and were asked to provide a justified recommendation to a ‘friend' (individual activity). Later, students were asked to consider how this recommendation might change if certain factors in the scenario changed (group discussion). Both these question styles were consistent with the format and scenario types from Question 10, 13 and 15 of the CAT.

Scenario: Melissa, a 34 year old social worker, is busy preparing for the birth of her second child. She is a client of the midwifery group practice program at the Women's and Children's Hospital and is scheduled to give birth in a private suite in 2 months’ time. Melissa will soon be on maternity leave, and she plans to take off as much time as she can to help her autistic son, Adrian, adjust to the new addition to the family. Even though she will be on leave, Melissa plans to keep Adrian in childcare, because she wants to minimise disruptions to his daily routine. However, his care is expensive and she may have to return to work sooner than desired, especially if her husband's plumbing business is slow. Melissa recently visited her local GP with Adrian for the last of his childhood vaccinations. When she was there, the GP asked her if she wanted to pre-book her baby in to start its vaccination schedule, as the Norwood-based clinic gets very busy in spring. Melissa said she wasn't sure if it was necessary to vaccinate her newborn against Hep B this time around. She is not sure when she will return to work, and she plans to breastfeed which she has been told confers immunity to the child. She is also a little apprehensive about the effectiveness and therefore the necessity of vaccinations because her son Adrian actually contracted whooping cough from childcare despite sticking to the recommended vaccine schedule.

a) As a trusted friend and known scientific thinker, Melissa has asked you for advice. Should she: -

* not vaccinate the new baby
* vaccinate the new baby as per schedule
* delay vaccination until ... (please specify when)

Download the information folder from FLO to help you decide which solution would be best for her situation. Explain your answer and provide at least two reasons for your decision.

b) Which information indicated risk?

c) Which information was just a distraction?

d) What happens to the risk level if we change the scenario?

- Melissa is offered an exciting opportunity with a new job working in a rural Aboriginal community
- Melissa is able to take advantage of a change in maternity leave payments and decides to take a full year of leave, removing her older son from childcare to spend as much time as possible with both children
- Melissa's partner is offered a new job teaching Plumbing Theory to apprentices at TAFE
- While on maternity leave Melissa is keen to stay active in her community and begins volunteering at the local animal shelter

---

Intervention task 3 - set as an individual thinking question

The final CT skill intervention occurred in week 11 and was again developed in line with CAT skill set 2. Students were presented with a scenario that required them to think through a real-world problem. Firstly, they were asked to suggest information that would be needed to make a life-changing decision, they were then presented with additional information and asked to make a recommendation. Finally, they were asked to identify and explain what changes to the scenario would cause them to alter/modify/switch their previous recommendation. This particular intervention was given at the start of class, with students working on the task individually, with the remainder of the class allocated to discussing the ethical issues related to the same theme but a different genetic condition.

Scenario: A couple underwent genetic testing and were told there is a 75% chance of their offspring developing a genetic disorder. They found this information quite overwhelming, and are now trying to decide if they will have children.

a) *What other factors should the couple consider when making this decision and why?*

The couple then sought genetic counselling. The counsellor was able to explain to them that the genetic disorder was beta-thalassemia, an inherited blood disorder, whose severity depends on the number of alleles (DNA) that are affected. In its minor form, the condition causes anaemia, however in its most severe form life expectancy is 20. The counsellor explained because that they were both carriers for the disease, the chances of their offspring developing some form of the condition were higher (75%) than if only one of them was a carrier (50%). The counsellor had them reflect on their own experiences with the condition. Only one reported having minimal symptoms, just a bit of tiredness and poor concentration. Since they had little experience with thalassemia at its worst, the counsellor referred them to some additional information.

- Symptoms: - https://www.nhlbi.nih.gov/health/health-topics/topics/thalassemia/signs
- Treatments: - https://www.nhlbi.nih.gov/health/health-topics/topics/thalassemia/treatment

Based on the information they received through the screening and counselling process, the couple have decided not to have children.

b) What changes to the circumstances might cause them to choose to have children and why?

APPENDIX K Additional results from Chapter 5: Exploring CT Development in Biology and Society

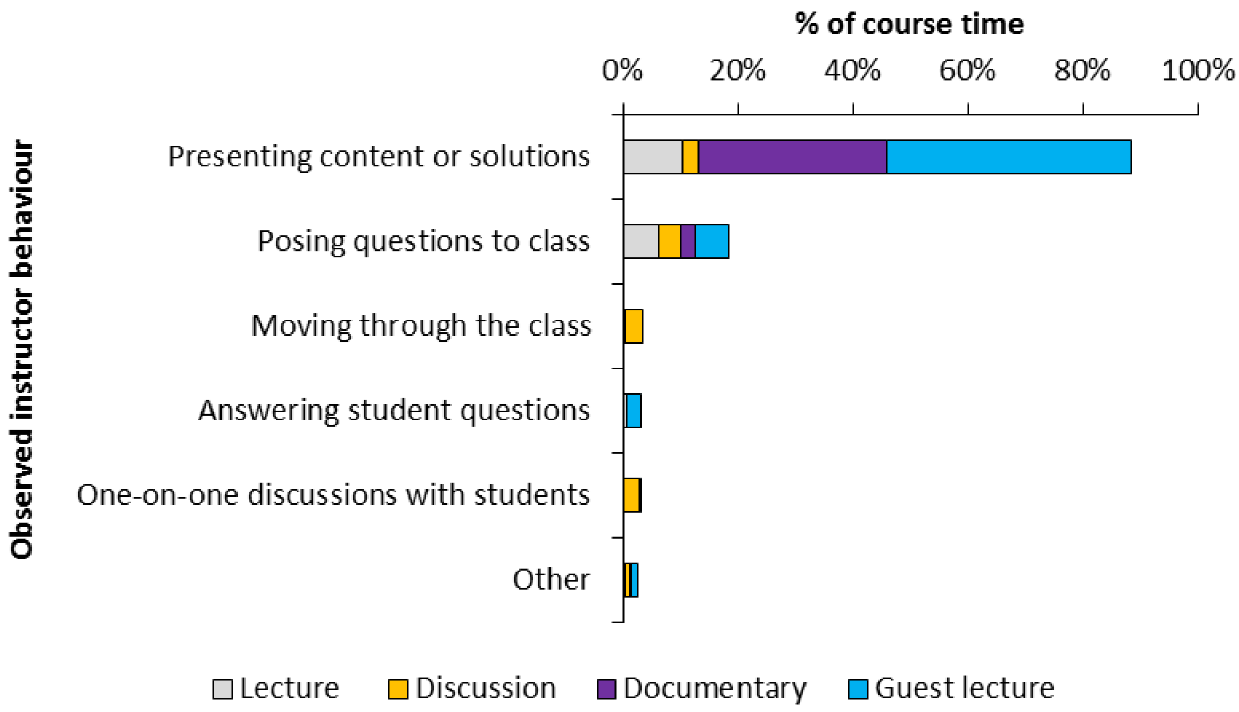


Figure K.1 Summary of how instructor time was spent in each component of in the B&S-V1 course.

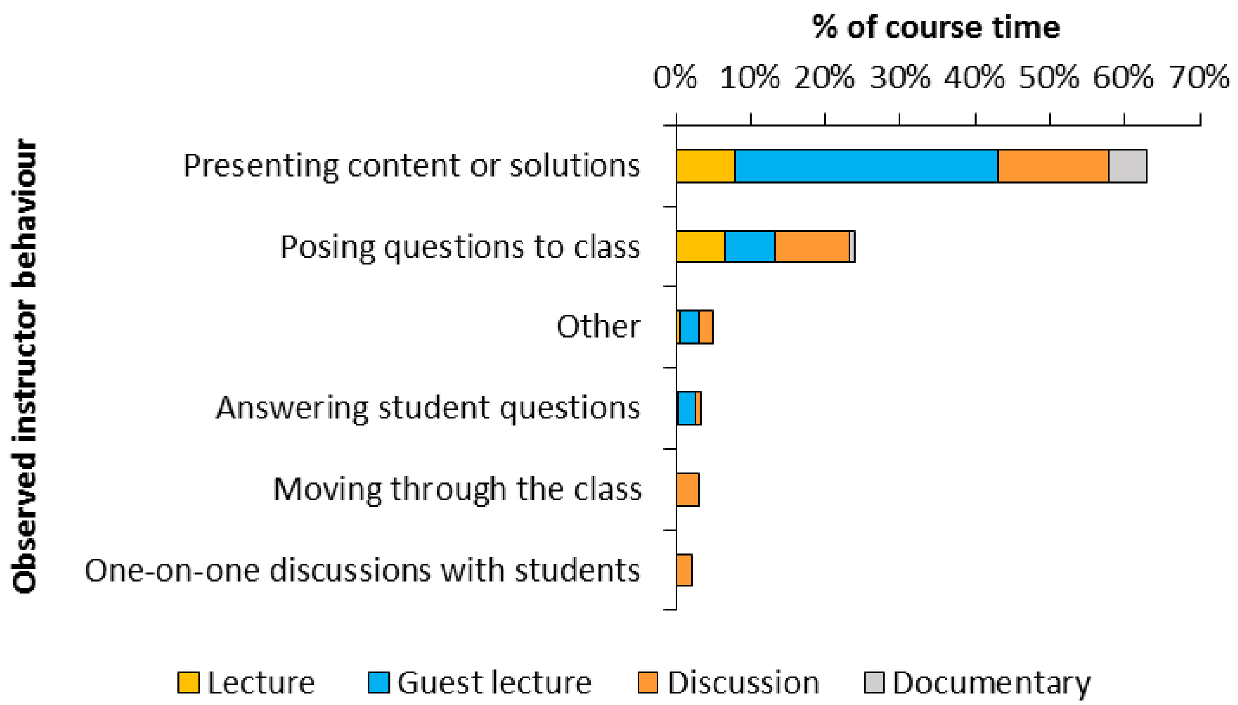


Figure K.2 Summary of how instructor time was spent in each component of in the B&S-V2 course.

Table K.1 Summary of the CAT results for B&S-V1 (n=30). Due to variations in the scales across questions, different statistical tests were applied as appropriate. These are indicated through the symbols near the question number and are explained below in the table notes.

	Q1ᶧ		Q2^		Q3^		Q4^	
	pre	post	pre	post	pre	post	pre	post
Maximum	1	1	3	3	3	3	4	4
Median	0	1	1	1	1	1	0	0
Mean	0.43	0.53	1.30	1.43	1.10	0.97	0.73	0.37
Sum	13	16	39	43	33	29	22	11
difference	3		4		-4		**-11**	
Statistical test statistic	0.6		28		-26		**-57**	
Statistical test: p-value§	0.438		0.595		0.533		**0.048**	

	Q5ᶧ		Q6^		Q7^		Q8ᶧ	
	pre	post	pre	post	pre	post	pre	post
Maximum	1	1	3	3	1	1	1	1
Median	1	1	2	2	0	0	1	1
Mean	0.60	0.63	1.57	1.47	0.30	0.07	0.60	0.70
Sum	18	19	47	44	9	2	18	21
difference	1		-3		**-7**		3	
Statistical test statistic	0.07		-12		**-35**		0.659	
Statistical test: p-value§	0.79		0.562		**0.039**		0.416	

	Q9^		Q10^		Q11^		Q12ᶧ	
	pre	post	pre	post	pre	post	pre	post
Maximum	2	2	4	4	2	2	1	1
Median	1	1	3	3	1	1	1	1
Mean	0.80	0.73	2.93	3.00	0.77	0.63	0.73	0.73
Sum	24	22	88	90	23	19	22	22
difference	-2		2		-4		0	
Statistical test statistic	-15		10		-30		0	
Statistical test: p-value§	0.79		0.859		0.424		0.999	

	Q13^		Q14^		Q15^		total#	
	pre	post	pre	post	pre	post	pre	post
Maximum	3	3	5	5	2	3	28	24
Median	1	1	3	3.5	0	0	16.5	16
Mean	1.00	1.00	2.37	2.63	0.50	0.30	15.73	15.20
Sum	30	30	71	79	15	9	472	456
difference	0		8		-6		-16	
Statistical test statistic	5		26		-28		0.777	
Statistical test: p-value§	0.929		0.459		0.185		0.443	

ᶧ chi-squared test statistic (chi-squared), with two-sided p-value. A contingency test comparing the proportions of 0 and 1 scores between pre and post.

^ Wilcoxon signed rank test statistic (*W*), with two-sided p-value. A non-parametric version of the paired t-test, to assess whether matched population mean ranks differ between pre and post.

paired t-test statistic (*t*), with two-sided p-value. A parametric test after confirmation of normality (using the Shapiro-Wilk and D'Agostino & Pearson normality tests) to assess whether population means differ pre and post.

§ In all cases, alpha was set to 0.05, where bolded p-values indicate where a significant difference lies.

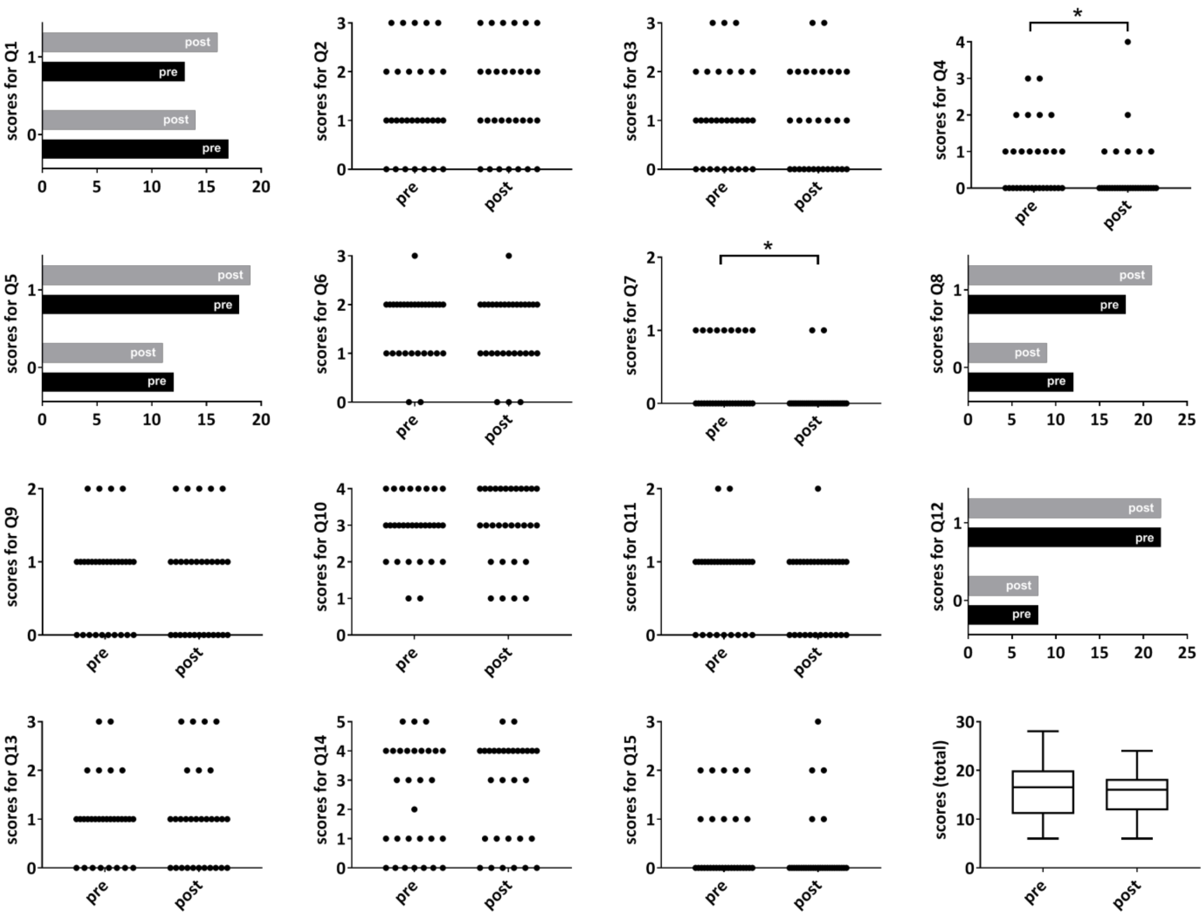


Figure K.3 Summary of the B&S-V1 cohort's pre-test and post-test CAT results (n=30). In all cases, alpha was set to 0.05, where an asterisk indicates a result where a significant difference lies and the number of asterisks denotes the level of significance, * <0.05, ** <0.01, * <0.001.**

Table K.2 Summary of the CAT results for B&S-V2 (n=37). Due to variations in the scales across questions, different statistical tests were applied as appropriate. These are indicated through the symbols near the question number and are explained below in the table notes.

	Q1□		Q2^		Q3^		Q4^	
	pre	post	pre	post	pre	post	pre	post
Maximum	1	1	3	3	3	3	4	4
Median	0	1	1	2	1	1	1	0
Mean	0.46	0.57	1.39	1.74	0.89	0.78	0.86	0.65
Sum	17	21	51.5	64.5	33	29	32	24
difference	4		**13**		-4		-8	
Statistical test statistic	0.865		**130**		-35		-57	
Statistical test: p-value§	0.352		**0.04**		0.504		0.308	

	Q5□		Q6^		Q7^		Q8□	
	pre	post	pre	post	pre	post	pre	post
Maximum	1	1	3	3	2	2	1	1
Median	1	1	1	2	0	0	1	1
Mean	0.73	0.81	1.32	1.57	0.32	0.22	0.65	0.70
Sum	27	30	49	58	12	8	24	26
difference	3		9		-4		2	
Statistical test statistic	0.687		99		-22		0.246	
Statistical test: p-value§	0.407		0.135		0.431		0.619	

	Q9^		Q10^		Q11^		Q12□	
	pre	post	pre	post	pre	post	pre	post
Maximum	2	2	4	4	2	2	1	1
Median	1	1	3	3	1	1	1	1
Mean	0.68	0.86	2.70	2.81	0.73	0.70	0.68	0.73
Sum	25	32	100	104	27	26	25	27
difference	7		4		-1		2	
Statistical test statistic	63		34		-8		0.258	
Statistical test: p-value§	0.143		0.606		0.954		0.611	

	Q13^		Q14^		Q15^		total#	
	pre	post	pre	post	pre	post	pre	post
Maximum	3	3	5	5	3	2	24	31
Median	1	1	1	1	0	0	13	14
Mean	0.89	1.03	1.41	1.81	0.24	0.19	13.96	15.18
Sum	33	38	52	67	9	7	516.5	561.5
difference	5		15		-2		45	
Statistical test statistic	50		76		-4		1.561	
Statistical test: p-value§	0.469		0.251		0.828		0.127	

□chi-squared test statistic (chi-squared), with two-sided p-value. A contingency test comparing the proportions of 0 and 1 scores between pre and post.

^ Wilcoxon signed rank test statistic (*W*), with two-sided p-value. A non-parametric version of the paired t-test, to assess whether matched population mean ranks differ between pre and post.

paired t-test statistic (*t*), with two-sided p-value. A parametric test after confirmation of normality (using the Shapiro-Wilk and D'Agostino & Pearson normality tests) to assess whether population means differ pre and post.

§ In all cases, alpha was set to 0.05, where bolded p-values indicate where a significant difference lies.

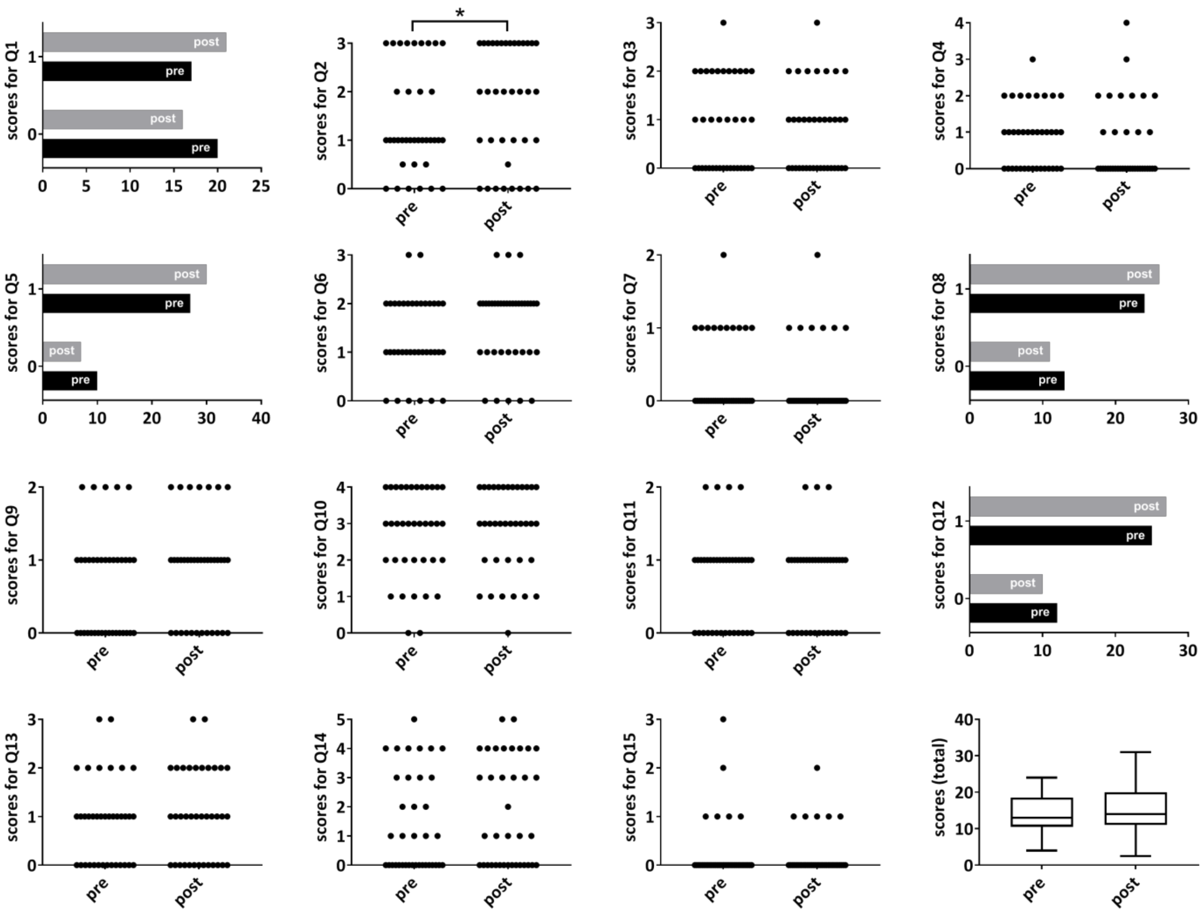


Figure K.4 Summary of the B&S-V2 cohort's pre-test and post-test CAT results (n=37). In all cases, alpha was set to 0.05, where an asterisk indicates a result where a significant difference lies and the number of asterisks denotes the level of significance, * <0.05, ** <0.01, * <0.001.**

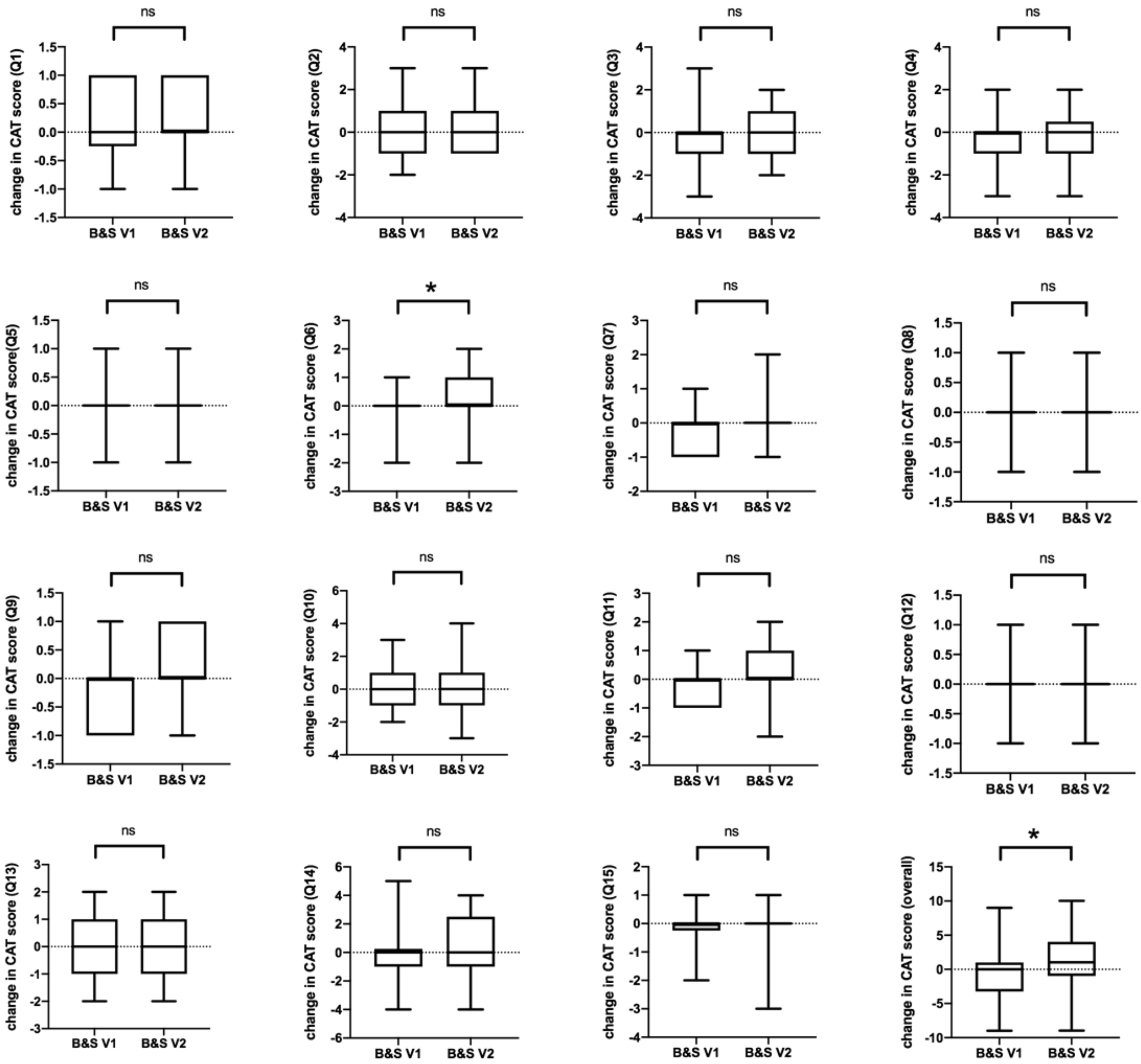


Figure K.5 Summary of the change in CAT performance (per question and overall difference) for version 1 and 2 of the B&S course. Significance explored using a Wilcoxon signed-rank test statistic (*W*), with a two-sided p=value. Alpha was set to 0.05, where the asterisk denote the level of significance, * <0.05, ** <0.01, * <0.001.**

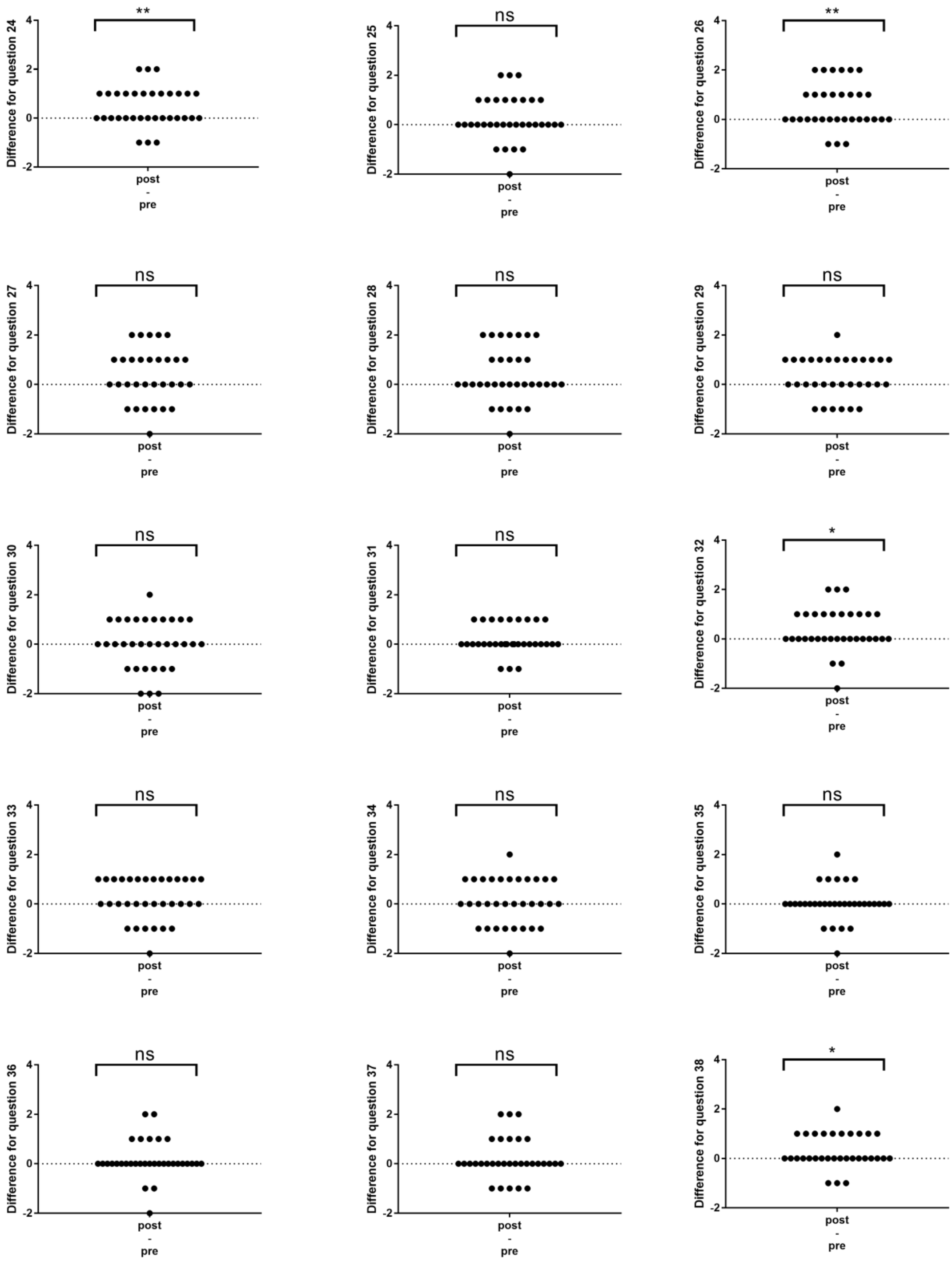


Figure K.6 Summary of the B&S-V2 student response differences for the disposition questions on the student perception survey. In all cases, alpha was set to 0.05, where an asterisk indicates a result where a significant difference lies and the number of asterisks denotes the level of significance, * <0.05, ** <0.01, * <0.001.**

Table K.3 B&S-V1 CAT report pre-test results compared to the national results for US freshman. ***Note.*** **Reprinted from Flinders University CAT Institution report (Centre for Assessment and Improvement of Learning, 2019b, p.14).**

Removed due to copyright restriction

Table K.4 B&S-V1 CAT report post-test results compared to the national results for US freshman. *Note.* Reprinted from Flinders University CAT Institution report (Centre for Assessment and Improvement of Learning, 2019b, p.29).

Removed due to copyright restriction

Table K.5 B&S-V2 CAT report pre-test results compared to the national results for US freshman. ***Note.*** **Reprinted from Flinders University CAT Institution report (Centre for Assessment and Improvement of Learning, 2019b, p.18).**

Removed due to copyright restriction

Table K.6 B&S-V2 CAT report post-test results compared to the national results for US freshman. *Note.* Reprinted from Flinders University CAT Institution report (Centre for Assessment and Improvement of Learning, 2019b, p.34).

Removed due to copyright restriction

BIBLIOGRAPHY

Ab Kadir, M. (2017). What teacher knowledge matters in effectively developing critical thinkers in the 21st century curriculum? *Thinking Skills and Creativity,* 23, 79-90. DOI:10.1016/j.tsc.2016.10.011

Ab Kadir, M. (2018). An inquiry into critical thinking in the Australian curriculum: examining its conceptual understandings and their implications on developing critical thinking as a “general capability” on teachers’ practice and knowledge. *Asia Pacific Journal of Education*, 38(4), 533-549. DOI:10.1080/02188791.2018.1535424

Abell, S. (2005). University science teachers as researchers: Blurring the scholarship boundaries. *Research in Science Education*, 35(2-3), 281–298. DOI:10.1007/s11165-004-5600-x

Abrami, P., Bernard, R., Borokhovski, E., Waddington, D., Wade, A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of Educational Research*, 85(2), 275-314. DOI:10.3102/0034654314551063

Abrami, P., Bernard, R., Borokhovski, E., Wade, A., Surkes, M., Tamim, R., & Zhang, D. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research*, 78(4), 1102–1134. DOI:10.3102/0034654308326084

Australian Curriculum, Assessment and Reporting Authority (2011) *General capabilities consultation report*. Sydney, New South Wales: ACARA. Retrieved from https://docs.acara.edu.au/resources/General_Capabilities_-_Consultation_Report_-_December__2011.pdf

Australian Curriculum, Assessment and Reporting Authority (2013). *The Australian curriculum: Senior Secondary Curriculum – Science*. Version 5.0. Retrieved from http://www.australiancurriculum.edu.au/Download/SeniorSecondary

Achen, R., & Lumpkin, A. (2015). Evaluating classroom time through systematic analysis and student feedback. *International Journal for the Scholarship of Teaching and Learning*, 9(2), Article 4, 1-8. DOI:10.20429/ijsotl.2015.090204

ACT CAAP (1989). *Science Test*. Iowa, ACT, Inc. Retrieved from http://www.act.org/caap/test/science.html

Adams, M., Whitlow, J., Stover, L., & Johnson, K. (1996). Critical thinking as an educational outcome: An evaluation of current tools of measurement. *Nurse Educator*, 21(3), 23-32. DOI:10.1097/00006223-199605000-00009

Addy, T., & Stevenson, M. (2014). Evaluating biological claims to enhance critical thinking skills through position statements. *Journal of Microbiology and Biology Education*, 15(1), 49-50. DOI:10.1128/jmbe.v15i1.663

Adler, J. (1991). Critical thinking, a deflated defense: A critical study of John E. McPeck's teaching critical thinking: Dialogue and dialectic. *Informal Logic*, 13(2), 61-78. DOI:10.22329/il.v13i2.2556

Adler, J. (1996). Teaching and the structural approach to fallacies. *Inquiry: Critical Thinking Across the Disciplines*, 15(4), 94-106. DOI:10.5840/inquiryctnews199615432

Akbulut, Y. (2015). Predictors of inconsistent responding in web surveys, *Internet Research*, 25(1), 131-147. DOI:10.1108/IntR-01-2014-0017

Akturk, A., & Sahin, I. (2011). Literature review on metacognition and its measurement. *Procedia Social and Behavioral Sciences*, 15, 3731-3736. DOI:10.1016/j.sbspro.2011.04.364

Alexander, B., Lindow, L., & Schock, M. (2008). Measuring the impact of cooperative learning exercises on student perceptions of peer-to-peer learning: a case study. *Journal of Physician Assistant Education*, 19(3), 18-25. DOI:10.1097/01367895-200819030-00005

Aliakbari, M., & Sadeghdaghighi, A. (2013). Teachers' perception of the barriers to critical thinking. *Procedia Social and Behavioral Sciences*, 70, 1-5. DOI:10.1016/j.sbspro.2013.01.031

Allen, M., Berkowitz, S., Hunt, S., & Louden, A. (1999). A meta-analysis of the impact of forensics and *Communication Education* on critical thinking. *Communication Education*, 48(1), 18-30. DOI:10.1080/03634529909379149

Almedia, L., & Franco, A. (2011). Critical thinking: Its relevance for education in a shifting society. *Revista de Psicología*, 29(1), 175-195. Retrieved from http://hdl.handle.net/1822/15482

Almeida, M., & Quintanilha, A. (2017). Of responsible research—Exploring the science-society dialogue in undergraduate training within the life sciences. *Biochemistry and Molecular Biology Education*, 45(1), 46-52. DOI:10.1002/bmb.20988

Aloisi, C., & Callaghan, A. (2018). Threats to the validity of the Collegiate Learning Assessment (CLA+) as a measure of critical thinking skills and implications for Learning Gain. *Higher Education Pedagogies*, 3(1), 57-82. DOI:10.1080/23752696.2018.1449128

Alwardai, M. (2014). *Islamic teachers perceptions of improving critical thinking skills in Saudi Arabian elementary schools* (Doctoral Dissertation, Southern Illinois University, Carbondale, Illinois). Retrieved from ProQuest Dissertations & Theses Global. (UMI 3640673).

Amin, A., & Adiansyah, R. (2018). Lecturers' perception on students' critical thinking skills development and problems faced by students in developing their critical thinking skills. *Jurnal Pendidikan Biologi Indonesia*, 4(1), 1-10. DOI:10.22219/jpbi.v4i1.5181.g5294

Amrein-Beardsley, A., & Popp, S. (2012). Peer observations among faculty in a college of education: investigating the summative and formative uses of the Reformed Teaching Observation Protocol (RTOP). *Educational Assessment, Evaluation and Accountability*, 24(1), 5-24. DOI:10.1007/s11092-011-9135-1

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature,* 567, 305-307. DOI:10.1038/d41586-019-00857-9

Anderson, L., Krathwohl, D., & Bloom, B. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Anderson, L., Sosniak, A., & Bloom, B. (1994). *Bloom's taxonomy: a forty-year retrospective.* Chicago: University of Chicago Press.

Anderson, M. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1), 32-46. DOI:10.1111/j.1442-9993.2001.01070.pp.x

Anderson, T. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). Wiley publications in statistics. New York: Wiley.

Andrews, R. (2009). *Argumentation in higher education: Improving practice through theory and research.* New York: Routledge.

Andrews, T., & Lemons, P. (2015). It's personal: Biology instructors prioritize personal evidence over empirical evidence in teaching decisions. *CBE-Life Sciences Education*, 14(1), ar7, 1-18. DOI:10.1187/cbe.14-05-0084

Andrews, T., Leonard, M., Colgrove, C., & Kalinowski, S. (2011). Active learning not associated with student learning in a random sample of college biology courses. *CBE-Life Sciences Education*, 10(4), 394–405. DOI:10.1187/cbe.11-07-0061

Angeli, C., & Valanides, N. (2009). Instructional effects on critical thinking: Performance on ill-defined issues. *Learning and Instruction,* 19(4), 322-334. DOI:10.1016/j.learninstruc.2008.06.010

Aretz, A., Bolen, M., & Devereux, K. (1997). Critical thinking assessment of college students, *Journal of College Reading and Learning,* 28(1), 12-23. DOI:10.1080/10790195.1997.10850050

Armitage, C., & Conner, M. (2001). Efficacy of the theory of planned behaviour: A meta-analytic review. *British Journal of Social Psychology*, 40(4), 471-499. DOI:10.1348/014466601164939

Arrington, W. (2017). *An analysis of the teaching of critical thinking skills in history in postsecondary education* (Doctoral dissertation, Texas Wesleyan University, Fort Wort, Texas). Retrieved from ProQuest Dissertations & Theses Global. (UMI 10636854).

Arter, J., & Salmon, J. (1987). *Assessing higher order thinking skills. A consumer's guide*. Portland, Oregon: Northwest Regional Educational Laboratory, Evaluation and Assessment.

Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses.* Chicago, Illinois: University of Chicago Press.

Arum, R., & Roksa, J. (2014). *Aspiring adults adrift: Tentative transitions of college graduates*. Chicago, Illinois: University of Chicago Press.

Arum, R., Cho, E., Kim, J., & Roksa, J. (2012). *Documenting uncertain times: Postgraduate transitions of the academically adrift cohort*. Brooklyn, New York: Social Science Research Council.

Arum, R., Roksa, J., & Cook, A. (Eds.). (2016). *Improving quality in American higher education: Learning outcomes and assessments for the 21st century.* San Francisco, California: Jossey-Bass Publishers.

Asimov, I., & Shulman, J. (Eds.). (1988). *Isaac Asimov's book of science and nature quotations*. New York: Weidenfeld & Nicolson.

Astuti, S., Suyanta, L., & Rohaeti, E. (2017). An integrated assessment instrument: Developing and validating instrument for facilitating critical thinking abilities and science process skills on electrolyte and nonelectrolyte solution matter. *AIP Conference Proceedings*, 1847(1), 050007, 1-6. DOI:10.1063/1.4983909

Atwa, S., Gauci-Mansour, V., Thomson, R., & Hegazi, I. (2019). Team-based and case-based learning: a hybrid pedagogy model enhancing students' academic performance and experiences at first-year tertiary level. *The Australian Educational Researcher*, 46(1), 93-112. DOI:10.1007/s13384-018-0282-y

Auerbach, A., & Andrews, T. (2018). Pedagogical knowledge for active-learning instruction in large undergraduate biology courses: a large-scale qualitative investigation of instructor thinking. *International Journal of STEM Education*, 5(1), 1-25. DOI:10.1186/s40594-018-0112-9

Australian Learning and Teaching Council (2011). *Resources to assist discipline communities to define Threshold Learning Outcomes (TLOs): An outcome of the ALTC's Learning and Teaching Academic Standards (LTAS) Project.* Strawberry Hills, New South Wales: Australian Learning and Teaching Council. Retrieved from http://www.olt.gov.au/resources

Badia, G. (2016). Question formation: A teachable art. *College & Undergraduate Libraries,* 23(2), 210-216. DOI:10.1080/10691316.2016.1167501

Baik, C., Naylor, R., Arkoudis, S., & Dabrowski, A. (2019). Examining the experiences of first-year students with low tertiary admission scores in Australian universities. *Studies in Higher Education*, 44(3), 526-538. DOI:10.1080/03075079.2017.1383376

Bailin, S. (2002). Critical thinking and science education. *Science and Education*, 11, 361–375. DOI:10.23/A:1016042608621

Bailin, S., & Battersby, M. (2017). What should I believe? Teaching critical thinking for reasoned judgment. *Teaching Philosophy,* 40(3), 275-295. DOI:10.5840/teachphil2017101672

Bailin, S., & Siegel, H. (2003). Critical thinking. In N. Blake, P. Smeyers, R. Smith & P. Standish (Eds.), *The Blackwell Guide to the philosophy of education* (pp. 181-193). Oxford: Blackwell Publishing (Holdings) Ltd.

Bailin, S., Case, R., Coombs, J., & Daniels, L. (1999). Conceptualizing critical thinking, *Journal of Curriculum Studies*, 31(3), 285-302. DOI:10.1080/002202799183133

Baillie, C., Jorre De St Jorre, T., & Hazel, E. (2017). *Improving teaching and learning in science and engineering laboratories*. Hammondville, New South Wales: Higher Education Research and Development Society of Australasia.

Ballen, C., Blum, J., Brownell, S., Hebert, S., Hewlett, J., Klein, J., McDonald, E., Monti, D., Nold, S., Slemmons, K., Soneral, P., & Cotner, S. (2017). A call to develop course-based undergraduate research experiences (CUREs) for nonmajors courses. *CBE—Life Sciences Education*, 16, mr2, 1-7. DOI:10.1187/cbe.16-12-0352

Bammer, G. (2013). *Disciplining interdisciplinarity: Integration and implementation sciences for researching complex real-world problems*. Canberra, Australia: Australian National University Press.

Barnett, J., & Francis, A. (2012). Using higher order thinking questions to foster critical thinking: a classroom study. *Educational Psychology,* 32(2), 201-211. DOI:10.1080/01443410.2011.638619

Barnett, S., & Ceci, S. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, 128(4), 612-637. DOI:10.1037/0033-2909.128.4.612

Baron, J. (1993). Why teach thinking? An essay. *Applied Psychology,* 42(3), 191-214.

Barrie, S. (2004). A research-based approach to generic graduate attributes policy, *Higher Education Research & Development,* 23(3), 261-275. DOI:10.1080/0729436042000235391

Barrie, S., Hughes, C., & Smith, C. (2009). *The national graduate attributes report: Integration and assessment of graduate attributes in curriculum*. Research Report. Sydney: Australian Learning and Teaching Council.

Bascandziev, I., & Harris, P. (2010). The role of testimony in young children's solution of a gravity-driven invisible displacement task. *Cognitive Development,* 25(3), 233-246. DOI:10.1016/j.cogdev.2010.06.002

Bassett, M. (2016). Teaching critical thinking without (much) writing: Multiple-choice and metacognition. *Teaching Theology & Religion*, 19(1), 20-40. DOI:10.1111/teth.12318

Bathgate, M., Aragón, O., Cavanagh, A., Frederick, J., & Graham, M. (2019a). Supports: A Key Factor in Faculty Implementation of Evidence-Based Teaching. *CBE—Life Sciences Education*, 18(2), ar22, 1-9. DOI:10.1187/cbe.17-12-0272

Bathgate, M., Aragón, O., Cavanagh, A., Waterhouse, J., Frederick, J., & Graham, M. (2019b). Perceived supports and evidence-based teaching in college STEM. *International Journal of STEM Education*, 6(11), 1-14. DOI:10.1186/s40594-019-0166-3

Beck, C., & Blumer, L. (2016). Alternative realities: Faculty and student perceptions of instructional practices in laboratory courses. *CBE-Life Sciences Education*, 15(4), ar52, 1-10. DOI:10.1187/cbe.16-03-0139

Beck, C., Butler, A., & Burke da Silva, K. (2014). Promoting inquiry-based teaching in laboratory courses: are we meeting the grade? *CBE-Life Sciences Education*, 13(3), 444-452. DOI:10.1187/cbe.13-12-0245

Beck, J. (2007). An exploration of the relationship between case study methodology and learning style preference. *Journal of Science Teacher Education*, 18(3), 423-430. DOI:10.1007/s10972-007-9056-5

Beghetto, R. (2002). Thinking about the outside of the box: A pedagogical framework for thinking critically about creativity. *Inquiry: Critical Thinking Across the Disciplines*, 21(2), 33-39. DOI:10.5840/inquiryctnews20022121

Behar-Horenstein, L., & Niu, L. (2011). Teaching critical thinking skills in higher education: A review of the literature. *Journal of College Teaching & Learning*, 8(2), 25-41. Retrieved from http://journals.cluteonline.com/index.php/TLC/article/view/3554

Bell, C., Dobbelaer, M., Klette, K., & Visscher, A. (2019). Qualities of classroom observation systems. *School Effectiveness and School Improvement*, 30(1), 3-29. DOI:10.1080/09243453.2018.1539014

Bellaera, L., Debney, L., & Baker, S. (2018). Subject Comprehension and Critical Thinking: An Intervention for Subject Comprehension and Critical Thinking in Mixed-Academic-Ability University Students. *The Journal of General Education,* 65(3-4), 264-282. DOI:10.5325/jgeneeduc.65.3-4.0264

Benjamin, R. (2012). *The seven red herrings about standardized assessments in higher education*. National Institute for Learning Outcomes Assessment Occasional Paper No. 15, Champaign, Illinois: National Institute for Learning Outcomes Assessment. Retrieved from http://learningoutcomesassessment.org/documents/HerringPaperFINAL.pdf

Benjamin, R., Klein, S., Steedle, J., Zahner, D., Elliot, S., & Patterson, J. (2013). The case for critical-thinking skills and performance assessment. *Council for Aid to Education Occasional Paper*, 1. New York: Council to Aid Education. Retrieved from http://cae.org/images/uploads/pdf/The_Case_for_Critical_Thinking_Skills.pdf

Bennett, J., Hogarth, S., & Lubben, F. (2003). A systematic review of the effects of context-based and Science – Technology - Society (STS) approaches in the teaching of secondary science. Version 1.1. *Research Evidence in Education Library*. London: EPPICentre, Social Science Research Unit, Institute of Education.

Bensley, D., Rainey, C., Murtagh, M., Flinn, J., Maschiocchi, C., Bernhardt, P., & Kuehne, S. (2016). Closing the assessment loop on critical thinking: The challenges of multidimensional testing and low test-taking motivation. *Thinking Skills and Creativity,* 21, 158-168. DOI:10.1016/j.tsc.2016.06.006

Bezanilla, M., Fernández-Nogueira, D., Poblete, M., & Galindo-Domínguez, H. (2019). Methodologies for teaching-learning critical thinking in higher education: The teacher's view. *Thinking Skills and Creativity*, 33, 100584, 1-10. DOI:10.1016/j.tsc.2019.100584

Bernstein, J. (2015). Transdisciplinarity: A review of its origins, development, and current issues**.** *Journal of Research Practice*, 11(1), Article R1, 1-20. Retrieved from http://jrp.icaap.org/index.php/jrp/article/view/510/412

Biggs, J., & Collis, K. (1982). *Evaluating the Quality of Learning: the SOLO taxonomy.* New York: Academic Press.

Birmingham, S. (2015). *Higher Education Standards Framework (Threshold Standards) 2015*. Canberra, Australian Capital Territory: Commonwealth of Australia. Retrieved from http://hdl.voced.edu.au/10707/381937.

Black, B. (2009, November). *Introducing a new subject and its assessment in schools: the challenges of introducing CT AS/A level in the UK*. Paper presented at the Association of Educational Assessment- Europe Conference. Retrieved from https://www.cambridgeassessment.org.uk/images/126122-introducing-a-new-subject-and-its-assessment-in-schools-the-challenges-of-introducing-critical-thinking-as-a-level-in-the-uk.pdf

Black, B. (2012). An overview of a programme of research to support the assessment of critical thinking. *Thinking Skills and Creativity,* 7(2), 122-133. DOI:10.1016/j.tsc.2012.04.003

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. A*ssessment in Education: Principles, Policy & Practice*, 25(6), 551-575. DOI:10.1080/0969594x.2018.1441807

Blaich, C., & Wise, K. (2011). *From Gathering to Using Assessment Results: Lessons from the Wabash National Study.* National Institute for Learning Outcomes Assessment Occasional Paper, No. 8. Urbana, Illinois: National Institute for Learning Outcomes Assessment, University of Illinois and Indiana University. Retrieved from www.learningoutcomesassessment.org/wpcontent/uploads/2019/02/OccasionalPaper8.pdf

Blaich, C., & Wise, K. (2017). Approaching big survey data one byte at a time. *New Directions for Student Services*, 159, 25-34. DOI:10.1002/ss.20224

Bloch, J., & Spataro, S. (2014). Cultivating critical-thinking dispositions throughout the business curriculum. *Business and Professional Communication Quarterly*, 77(3), 249-265. DOI:10.1177/2329490614538094

Bloom, B. (1994). Reflections on the development and use of the taxonomy. In L. Anderson, L. Sosniak & B. Bloom (Eds.), *Bloom's Taxonomy: A Forty-Year Retrospective* (pp. 1-8). Chicago: National Society for the Study of Education.

Bloom, B., Englehart, M., Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain.* New York: Longmans, Green and Co.

Blumer, L., & Beck, C. (2019). Laboratory courses with guided-inquiry modules improve scientific reasoning and experimental design skills for the least-prepared undergraduate students. *CBE-Life Sciences Education*, 18(1), ar2, 1-13. DOI:10.1187/cbe.18-08-0152

Bonwell, C., & Eison, J. (1991). *Active learning: Creating excitement in the classroom. 1991 ASHE-ERIC Higher Education Reports*. The George Washington University, Washington, DC: ERIC Clearinghouse on Higher Education.

Boon, M., & Van Baalen, S. (2019). Epistemology for interdisciplinary research–shifting philosophical paradigms of science. *European journal for philosophy of science*, 9(16), 1-28. DOI:10.1007/s13194-018-0242-4

Boonphadung, S. (2017). Critical thinking development: A comparison between the efficiency of mixed technique and a series of mini-lecture, KWL/TPS and presentation. *Journal of Advances in Humanities and Social Sciences,* 3(2), 95-104. DOI:10.20474/jahss-3.2.3

Boorstin, D. (1970). *A Case of Hypochondria*, Newsweek (6 July 1970).

Boorstin, D. (1983). *The Discoverers: A History of Man's Search to Know His World and Himself*, New York: Random House.

Borkowski, J. (1996). Metacognition: Theory or Chapter Heading? *Learning and Individual Differences,* 8(4), 391-402. DOI:10.1016/S1041-6080(96)90025-4

Borrego, M., & Henderson, C. (2014). Increasing the use of evidence-based teaching in STEM higher education: A comparison of eight change strategies. *Journal of Engineering Education,* 103(2), 220–252. DOI:10.1002/jee.20040

Bowman, N. (2010). College diversity experiences and cognitive development: a meta-analysis. *Review of Educational Research*, 80(1), 4–33. DOI:10.3102/0034654309352495

Bowman, N. (2014). The meaning and interpretation of college student self-reported gains. *New Directions for Institutional Research*, 2014(161), 59-68. DOI:10.1002/ir.20067

Box, C. (2019). *Formative assessment in United States classrooms: Changing the landscape of teaching and learning*. Cham, Switzerland: Palgrave Macmillan.

Bozzuto, D. (2017). *The utilization of inquiry-based science instruction in Connecticut* (Western Connecticut State University, Danbury, Connecticut). Retrieved from ProQuest Dissertations & Theses Global. (UMI 10286533).

Brabeck, M. (1983). Critical thinking skills and reflective judgment development: Redefining the aims of higher education. *Journal of Applied Developmental Psychology,* 4(1), 23-34. DOI:10.1016/0193-3973(83)90056-4

Bray, J., & Curtis, J. (1957). An ordination of upland forest communities of southern Wisconsin. *Ecological Monographs*, 27, 325-349. DOI:10.2307/1942268

Brookfield, S. (1987). *Developing Critical Thinkers*. Milton Keynes: Open University Press.

Brown, A., Bransford, J., Ferrara, R., & Campione, J. (1982). *Learning, remembering, and understanding*. Technical Report No. 244. Center for the Study of Reading, University of Illinois: Bolt, Beranek and Newman Inc. Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/17511/ctrstreadtechrepv01982i00244_opt.pdf

Brown, G., Andrade, H., & Chen, F. (2015). Accuracy in student self-assessment: directions and cautions for research. *Assessment in Education: Principles, Policy & Practice*, 22(4), 444-457. DOI:10.1080/0969594X.2014.996523

Brown, J. (2012). The current status of STEM education research. *Journal of STEM Education: Innovations and Research,* 13(5), 7-11. Retrieved from https://search.proquest.com/docview/1266493670?accountid=10910

Brownell, S., & Tanner, K. (2012). Barriers to Faculty Pedagogical Change: Lack of Training, Time, Incentives, and…Tensions with Professional Identity? *CBE-Life Sciences Education*, 11(4), 339–346. DOI:10.1187/cbe.12-09-0163

Brownell, S., Hekmat-Scafe, D., Singla, V., Seawell, P., Imam, J., Eddy, S., Stearns, T., & Cyert, M. (2015). A high-enrollment course-based undergraduate research experience improves student conceptions of scientific thinking and ability to interpret data. *CBE-Life Sciences Education*, 14(2), ar21, 1-14. DOI:10.1187/cbe.14-05-0092

Brownstein, N., Louis, T., O'Hagan, A., & Pendergast, J. (2019). The Role of Expert Judgment in Statistical Inference and Evidence-Based Decision-Making. *The American Statistician*, 73. DOI:10.1080/00031305.2018.1529623

Butler, A. (2018). *Critically thinking about critical thinking.* Lecture notes prepared for BIOL1112: Biology and Society, Flinders University, Bedford Park, 24 July.

Butler, A. (2019). *Critically thinking about critical thinking in science education: Reflecting on the perceptions and actions of Australian senior secondary and tertiary educators.* (Unpublished doctoral dissertation). Flinders University, Bedford Park, South Australia.

Butler, H., Dwyer, C., Hogan, M., Franco, A., Rivas, S., Saiz, C., & Almeida, L. (2012). The Halpern Critical Thinking Assessment and real-world outcomes: Cross-national applications. *Thinking Skills and Creativity,* 7(2), 112-121. DOI:10.1016/j.tsc.2012.04.001

Buxner, S., Impey, C., Romine, J., & Nieberding, M. (2018). Linking introductory astronomy students' basic science knowledge, beliefs, attitudes, sources of information, and information literacy. *Physical Review Physics Education Research*, 14(1), 010142, 1-17. DOI:10.1103/PhysRevPhysEducRes.14.010142

Byrne, M., & Johnstone, A. (1987). Critical thinking and science education. *Studies in Higher Education,* 12(3), 325–339. DOI:10.1080/03075078712331378102

Case, D., & Given, L. (2016). *Looking for information: A survey of research on information seeking, needs, and behavior* (4th ed.). Bingley, United Kingdom: Emerald Group Publishing.

Campbell, C., Cabrera, A., Michel, J., & Patel, S. (2017). From comprehensive to singular: A latent class analysis of college teaching practices. *Research in Higher Education*, 58(6), 581–604. DOI:10.1007/s11162-016-9440-0

Cao, T. H., Jung, J., & Lee, J. (2017). Assessment in gifted education: A review of the literature from 2005 to 2016. *Journal of Advanced Academics*, 28(3), 163-203. DOI:10.1177/1932202X17714572

Capossela, T. (1998). What is critical writing? In T. Capossela (Ed.), *The critical writing workshop: Designing writing assignments to foster critical thinking* (pp. 1-10). Portsmouth, New Hampshire: Heinemann.

Carbone, A., Drew, S., Ross, B., Ye, J., Phelan, L., Lindsay, K., & Cottman, C. (2019). A collegial quality development process for identifying and addressing barriers to improving teaching. *Higher Education Research & Development*, 39(7), 1356-1370. DOI:10.1080/07294360.2019.1645644

Cargas, S., Williams, S., & Rosenberg, M. (2017). An approach to teaching critical thinking across disciplines using performance tasks with a common rubric. *Thinking Skills and Creativity,* 26, 24-37. DOI:10.1016/j.tsc.2017.05.005

Carmel, J., & Yezierski, E. (2013). Are we keeping the promise? Investigation of students' critical thinking growth. *Journal of College Science Teaching,* 42(5), 71-81. Retrieved from https://www.jstor.org/stable/43631585

Carson, S. (2015). Targeting critical thinking skills in a first-year undergraduate research course. *Journal of Microbiology & Biology Education,* 16(2), 148. DOI:10.1128/jmbe.v16i2.935

Caspersen, J., Smeby, J., & Olaf Aamodt, P. (2017). Measuring learning outcomes. *European Journal of Education,* 52(1), 20-30. DOI:10.1111/ejed.12205

Celuch, K., Kozlenkova, I., & Black, G. (2010). An exploration of self-efficacy as a mediator of skill beliefs and student self-identity as a critical thinker. *Marketing Education Review,* 20(3), 70-80. DOI:10.2753/MER1052-8008200306

Center for Assessment and Improvement of Learning (2013). *Critical thinking Assessment Test training manual*, Version 8. Cookeville, Tennessee: Tennesse Tech University.

Centre for Assessment and Improvement of Learning (2019a). *CAT App PowerPoint*. Retrieved from https://www.tntech.edu/cat/pdf/training/CAT_App_Presentation.pdf

Centre for Assessment and Improvement of Learning (2019b). *Critical thinking Assessment Test institution report for Flinders University*. Cookeville, Tennessee: Tennesse Tech University.

Cetto, A., Schneegans S., & Moore, H. (Eds.). (2000). *World Conference on Science*: *Science for the twenty-first century: A new commitment - Declaration on science and the use of scientific knowledge*. Paris: United Nations Educational, Scientific and Cultural Organization. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000122938

Chabalengula, V., & Mumba, F. (2012). Inquiry-based science education: A scenario on Zambia's high school science curriculum. *Science Education International,* 23(4), 307-327. Retrieved from https://files.eric.ed.gov/fulltext/EJ1001626.pdf

Chan, C., Fong, E., Luk, L., & Ho, R. (2017). A review of literature on challenges in the development and implementation of generic competencies in higher education curriculum. *International Journal of Educational Development*, 57, 1-10. DOI:10.1016/j.ijedudev.2017.08.010

Chaplin, S. (2009). Assessment of the impact of case studies on student learning gains in an introductory biology course. *Journal of College Science Teaching,* 39(1), 72-79. Retrieved from https://search.proquest.com/docview/200374187?accountid=10910

Choy, S., & Cheah, P. (2009). Teacher perceptions of critical thinking among students and its influence on higher education. *International Journal of Teaching and Learning in Higher Education,* 20(2), 198-206. Retrieved from https://files.eric.ed.gov/fulltext/EJ864337.pdf

Chung, H., & Behan, K. (2010). Peer sharing facilitates the effect of inquiry-based projects on science learning. *The American Biology Teacher,* 72(1), 24-29. DOI:10.1525/abt.2010.72.1.7

Cian, H., Marshall, J., & Qian, M. (2018). Inquiry classroom patterns of student cognitive engagement: An analysis using growth curve modeling. *Journal of Science Teacher Education*, 29(4), 326-346. DOI:10.1080/1046560X.2018.1456884

Clarke, K., & Gorley, R. (2006). *PRIMER v6: User Manual/Tutorial*. Plymouth, Devon: PRIMER-E Limited.

Clarke, K., & Warwick, R. (2001). *Change in marine communities: an approach to statistical analysis and interpretation* (2nd ed.). Plymouth, Devon: PRIMER-E Limited.

Claxton, G. (2012). Turning thinking on its head: How bodies make up their minds. *Thinking Skills and Creativity,* 7(2), 78-84. DOI:10.1016/j.tsc.2012.03.004

Clayton, L. (2006). Concept mapping: an effective, active teaching-learning method. *Nursing Education Perspectives*, 27(4), 197-203. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/16921805

Cleveland, L., Olimpo, J., & DeChenne-Peters, S. (2017). Investigating the relationship between instructors' use of active-learning strategies and students' conceptual understanding and affective changes in introductory biology: A comparison of two active-learning environments. *CBE—Life Sciences Education*, 16(2), ar19, 1-10. DOI:10.1187/cbe.16-06-0181

Clifford, J., Boufal, M., & Kurtz, J. (2004). Personality Traits and Critical thinking Skills in College Students: Empirical Test of a Two-Factor Theory. *Assessment*, 11(2), 169-176. DOI:10.1177/1073191104263250

Clyde, F. (2004). Can case studies be used to teach critical thinking? *Journal of College Science Teaching,* 33(6), 12-14. Retrieved from https://www.jstor.org/stable/26491296

Coe, R. (2013). *Improving education: A triumph of hope over experience.* Ushaw College, Durham: Centre for Evaluation and Monitoring, Durhman University. Retrieved from http://www.cem.org/attachments/publications/ImprovingEducation2013.pdf

Coker, P. (2009). *The effects of an experiential learning program on the clinical reasoning and critical thinking skills of occupational therapy students* (Doctoral dissertation, Capella University, Minneapolis, Minnesota). Retrieved from ProQuest Dissertations & Theses Global. (UMI 3344517).

Cole, D., Ullman, J., Gannon, S., & Rooney, P. (2015). Critical thinking skills in the International Baccalaureate's "Theory of Knowledge" subject: Findings from an Australian study. *Australian Journal of Education,* 59(3), 247-264. DOI:10.1177/0004944115603529

Collaço, C. (2017). Increasing Student Engagement in Higher Education. *Journal of Higher Education Theory and Practice,* 17(4), 40-47. DOI:10.33423/jhetp.v17i4.1545

Collins, E., & Calhoun, T. (2014). Raising the bar in freshman science education: Student lectures, scientific papers, and independent experiments. *Journal of College Science Teaching,* 43(4), 26-35. Retrieved from https://www.jstor.org/stable/43632009

Correia, V. (2018). Contextual debiasing and critical thinking: Reasons for optimism. *Topoi*, 37(1), 103-111. DOI:10.1007/s11245-016-9388-x

Costandius, E., Blackie, M., Leibowitz, B., Nell, I., Rhoda, M., Rosochacki, S., & Young, G. (2015). Exploring notions of critical citizenship. In M. Davies, M and R. Barnett (Eds.), *The Palgrave handbook of critical thinking in higher education* (pp. 545-558). New York: Palgrave Macmilian.

Cotner, S., Thompson, S., & Wright, R. (2017). Do biology majors really differ from non–STEM majors? *CBE—Life Sciences Education*, 16(3), ar48, 1-8. DOI:10.1187/cbe.16-11-0329.

Cowden, C., & Santiago, M. (2016). Interdisciplinary explorations: Promoting critical thinking via problem-based learning in an advanced biochemistry class. *Journal of Chemical Education*, 93, 464-469. DOI:10.1021/acs.jchemed.5b00378

Cross, D., & Paris, S. (1988). Developmental and instructional analyses of children's metacognition and reading comprehension. *Journal of Educational Psychology,* 80(2), 131-142. DOI:10.1037/0022-0663.80.2.131

Csapó, B., Ainley, R., Bennett, R., Latour, T., & Law, N. (2012). Chapter 4: Technological issues for computer based assessment. In P. Griffin, B. McGaw & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 143-230). New York: Springer.

Cuban, L. (1984). Policy and research dilemmas in the teaching of reasoning: unplanned designs. *Review of Educational Research*, 54(4), 655-681. DOI:10.3102/00346543054004655

Culver, K., Braxton, J., & Pascarella, E. (2019). Does teaching rigorously really enhance undergraduates' intellectual development? The relationship of academic rigor with critical thinking skills and lifelong learning motivations. *Higher Education*, 78(4), 611–627. DOI:10.1007/s10734-019-00361-z

Czajka, C., & McConnell, D. (2019). The adoption of student-centered teaching materials as a professional development experience for college faculty. *International Journal of Science Education*, 41(5), 693-711. DOI:10.1080/09500693.2019.1578908

Danczak, S., Thompson, C., & Overton, T. (2017). 'What does the term Critical Thinking mean to you?' A qualitative analysis of chemistry undergraduate, teaching staff and employers' views of critical thinking. *Chemistry Education Research and Practice*, 18(3), 420-434. DOI:10.1039/C6RP00249H

Danvers, E. (2015). Criticality's affective entanglements: rethinking emotion and critical thinking in higher education. *Gender and Education*, 28(2), 282–297. DOI:10.1080/09540253.2015.1115469

Davidson, J., & Sternberg, R. (1998). Smart problem solving: How metacognition helps. In D. Hacker, J. Dunlosky & A. Graesser (Eds.), *Metacognition in theory and practice* (pp. 47-68). Mahwah, New Jersey: Erlbaum.

Davies, M., & Barnett, R. (2015). *The Palgrave handbook of critical thinking in higher education.* New York: Palgrave Macmilian.

Davies, W. (2006). An 'infusion' approach to critical thinking: Moore on the critical thinking debate. *Higher Education Research and Development,* 25(2), 179–193. DOI:10.1080/07294360600610420

Davis, J., Clayton, C., & Broome, J. (2017). Thinking like researchers: action research and its impact on novice teachers' thinking. *Educational Action Research,* 26(1), 59–74. DOI:10.1080/09650792.2017.1284012

de Bie, H., Wilhelm, P., & van der Meij, H. (2015). The Halpern Critical Thinking Assessment: Toward a Dutch appraisal of critical thinking. *Thinking Skills and Creativity,* 17, 33-44. DOI:10.1016/j.tsc.2015.04.001

De Bono, E. (1984). Critical thinking is not enough. *Educational Leadership*, 42(1), 16-17. Retrieved from http://www.ascd.org/publications/educational-leadership/sept84/vol42/num01/toc.aspx

de Kraker-Pauw, E., van Wesel, F., Verwijmeren, T., Denessen, E., & Krabbendam, L. (2016). Are teacher beliefs gender-related? *Learning and Individual Differences,* 51, 333-340. DOI:10.1016/j.lindif.2016.08.040

Delić, H., & Bećirović, S. (2016). Socratic method as an approach to teaching. *European Researcher, Series A,* 111(10), 511-517. DOI:10.13187/er.2016.111.511

Desimone, L., & Le Floch, K. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis,* 26(1), 1-22. DOI:10.3102/01623737026001001

Detlor, B., Booker, L., Serenko, A., & Julien, H. (2012). Student perceptions of information literacy instruction: The importance of active learning. *Education for Information*, 29(2), 147–161. DOI:10.3233/efi-2012-0924

Dewey, J. (1910). *How we think*. Boston: D.C. Heath & Co. Publishers.

Dewey, J. (1929). *The quest for certainty: a study of the relation of knowledge and action.* New York: Minton, Balch and Company.

Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process.* Boston: D. C. Heath and Company.

Dewey, J. (1986). Experience and education. *The Educational Forum,* 50(3), 241-252. DOI:10.1080/00131728609335764

Dignath, C., & Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning,* 3, 231-264. DOI:10.1007/s11409-008-9029-x

Dilley, A., Kaufman, J., Kennedy, C., & Plucker, J. (2015). *What we know about critical thinking.* Washington, DC: Partnership for 21st Century Skills.

Dillon, H., James, C., Prestholdt, T., Peterson, V., Salomone, S., & Anctil, E. (2019). Development of a formative peer observation protocol for STEM faculty reflection. *Assessment & Evaluation in Higher Education*, Advanced online publication, 1–14. DOI:10.1080/02602938.2019.1645091

Dinsmore, D., Alexander, P., & Loughlin, S. (2008). Focusing the conceptual lens on metacognition, self-regulation and self-regulated learning. *Educational Psychology Review,* 20(4), 391-409. DOI:10.1007/s10648-008-9083-6

Doeke, E., & Maire, Q. (2019). Key skills for the 21st century: An evidence-based review. In K. Cockel (Ed.), *Proceedings and Program Research Conference 2019: Preparing students for life in the 21st century - Identifying, developing and assessing what matters* (pp. 73-76). Camberwell, VIC: Australian Council for Educational Research. Retrieved from https://www.acer.org/files/Research-Conference-2019-Proceedings-and-Program.pdf

Doherty, J., Hansen, M., & Kaya, K. (1999). Teaching Information Skills in the Information Age: The Need for Critical Thinking. *Library Philosophy and Practice*, 1(2), 1-9. Retrieved from http://www.webpages.uidaho.edu/~mbolin/doherty.pdf

Dolan, E. (2015). Biology education research 2.0. *CBE-Life Sciences Education*, 14(4), ed1, 1-2. DOI:10.1187/cbe.15-11-0229

Dolan, E., Elliott, S., Henderson, C., Curran-Everett, D., St. John, K., & Ortiz, P. (2018). Evaluating Discipline-Based Education Research for promotion and tenure. *Innovative Higher Education,* 43(1), 31–39. DOI:10.1007/s10755-017-9406-y

Donleavy, G. (2012). Proclaimed graduate attributes of Australian universities: patterns, problems and prospects. *Quality Assurance in Education*, 20(4), 341 – 356. DOI:10.1108/09684881211263984

Douglas, E. (2012). Defining and measuring critical thinking in engineering. *Procedia Social and Behavioral Sciences*, 56, 153-159. DOI:10.1016/j.sbspro.2012.09.642

Downs-Lombardi, J. (1995). Teaching styles that encourage and discourage thinking skills. *Inquiry: Critical Thinking Across the Disciplines*. 15(2), 67-71. DOI:10.5840/inquiryctnews199515227

Dumitru, D. (2012). Critical thinking and integrated programs. The problems of transferability. *Procedia Social and Behavioral Sciences*, 33, 143-147. DOI:10.1016/j.sbspro.2012.01.100

Dunne, G. (2015). Beyond critical thinking to critical being: Criticality in higher education and life. *International Journal of Education Research,* 71, 86-99. DOI:10.1016/j.ijer.2015.03.003

Duron, R., Limbach, B., & Waugh, W. (2006). Critical thinking framework for any discipline. *International Journal of Teaching and Learning in Higher Education,* 17(2), 160-166. Retrieved from http://www.isetl.org/ijtlhe/pdf/IJTLHE55.pdf

Dweck, C., Walton, G., & Cohen, G. (2014). *Academic tenacity: Mindsets and skills that promote long-term learning.* Seattle, Washington: Bill and Melinda Gates Foundation.

Dwyer, C. (2017). *Critical thinking: Conceptual perspectives and practical guidelines*. Cambridge: Cambridge University Press.

Dwyer, C., & Eigenauer, J. (2017). To teach or not to teach critical thinking: A reply to Huber and Kuncel. *Thinking Skills and Creativity,* 26, 92-95. DOI:10.1016/j.tsc.2017.08.002

Dwyer, C., Hogan, M., & Stewart, I. (2014). An integrated critical thinking framework for the twenty-first century. *Thinking Skills and Creativity,* 12, 43-52. DOI:10.1016/j.tsc.2013.12.004

Dwyer, C., Hogan, M., Harney, O., & Kavanagh, C. (2017). Facilitating a student-educator conceptual model of dispositions towards critical thinking through interactive management. *Educational Technology Research and Development*, 65(1), 47-73. DOI:10.1007/s11423-016-9460-7

Eastman, M. (2017). *The journey from engineering educator to engineering education researcher* (Doctoral dissertation, State University of New York, Buffalo, New York). Retrieved from ProQuest Dissertations & Theses Global. (UMI 10279363).

Ebert-May, D., Derting, T., Hodder, J., Momsen, J., Long, T., & Jardeleza S. (2011). What we say is not what we do: Effective evaluation of faculty professional development programs. *BioScience,* 61, 550–558. DOI:10.1525/bio.2011.61.7.9

Eddy, S., & Hogan, K. (2014). Getting under the hood: How and for whom does increasing course structure work? *CBE—Life Sciences Education*, 13(3), 453-468. DOI:10.1187/cbe.14-03-0050

Eddy, S., Converse, M., & Wenderoth, M. (2015). PORTAAL: a classroom observation tool assessing evidence-based teaching practices for active learning in large science, technology, engineering, and mathematics classes. *CBE-Life Sciences Education*, 14(2), ar23, 1-16. DOI:10.1187/cbe.14-06-0095

Edwards, L. (2017). The craft of infusing critical thinking skills: A mixed method research on implementation and student outcome. *Journal on Centers for Teaching and Learning,* 9, 47-72. Retrieved from http://openjournal.lib.miamioh.edu/index.php/jctl/article/view/189

Edwards, S. (2007). Critical thinking: A two-phase framework. *Nurse Education in Practice,* 7, 303-314. DOI:10.1016/j.nepr.2006.09.004

Ellerman, C., Kataoka-Yahhiro, M., & Wong, L. (2006). Logic modes used to enhance critical thinking. *Journal of Nursing Education,* 45(6), 229-227. Retrieved from https://search.proquest.com/docview/203927585?accountid=10910

Ellerton, P. (2011). Reason to think. *Issues*, 95, 33-35. Retrieved from https://search.informit.com.au/fullText;res=IELAPA;dn=201107098

Ellerton, P. (2015). Metacognition and critical thinking: some pedagogical imperatives. In Davies, M., & Barnett, R. (Eds.), *Palgrave handbook of critical thinking in higher education*, 409-426. New York: Palgrave Macmillan. DOI:10.1007/978-1-137-37805-7_25

Ellerton, P. (2017). Pragmatic epistemology, inquiry values and education for thinking. In M. Gregory, J. Haynes & K. Murris (Eds.), *The Routledge International Handbook of philosophy for children* (pp. 111-118). New York: Routledge.

Elliott, K., Boin, A., Irving, H., Johnson, E., & Galea, V. (2010a). *Teaching scientific inquiry skills: A handbook for bioscience educators in Australian universities.* Strawberry Hills, New South Wales: Australian Learning and Teaching Council.

Elliott, K., Boin, A., Irving, H., Johnson, E., & Galea, V. (2010b). *Enhancing the learning of scientific inquiry skills for bioscience students in Australian universities: Final report.* Strawberry Hills, New South Wales: Australian Learning and Teaching Council.

Elliott, R., & Paton, V. (2018). US higher education reform: Origins and impact of student curricular choice. *International Journal of Educational Development*, 61, 1-4. DOI:10.1016/j.ijedudev.2017.11.008

Engle, R., Lam, D., Meyer, X., & Nix, S. (2012). How does expansive framing promote transfer? Several proposed explanations and a research agenda for investigating them. *Educational Psychologist,* 47(3), 215-231. DOI:10.1080/00461520.2012.695678

Ennis, R. (1962). A concept of critical thinking. *Harvard Educational Review*, 32(1), 81-111. Retrieved from ERIC (ED021832).

Ennis, R. (1985). A logical basis for measuring critical thinking skills. *Educational Leadership*, 43(2), 44–48. Retrieved from https://pdfs.semanticscholar.org/80a7/c7d4a98987590751df4b1bd9adf747fd7aaa.pdf

Ennis, R. (1989). Critical thinking and subject specificity: Clarification and needed research. *Educational Researcher,* 18(3), 4-10. DOI:10.3102/0013189X018003004

Ennis, R. (1990). The extent to which critical thinking is subject-specific: Further clarification. *Educational Researcher,* 19(4), 13-16. DOI:10.3102/0013189X019004013

Ennis, R. (1991). Critical thinking: A streamlined conception. *Teaching Philosophy,* 14(1), 5-25. DOI:10.5840/teachphil19911412

Ennis, R. (1997). Incorporating critical thinking in the curriculum: An introduction to some basic issues. *Inquiry: Critical Thinking Across the Disciplines*, 16(3), 1-9. DOI:10.5840/inquiryctnews199716312

Ennis, R. (2011a). Critical thinking: Reflection and perspective, Part 1. *Inquiry: Critical Thinking Across the Disciplines*, 26(1), 4-18. DOI:10.5840/inquiryctnews20112613

Ennis, R. (2011b). Critical thinking: Reflection and perspective, Part 2. *Inquiry: Critical Thinking Across the Disciplines*, 26(2), 5-19. DOI:10.5840/inquiryctnews201126215

Ennis, R. (2013). *Twenty-one strategies and tactics for teaching critical thinking*. Retrieved from http://www.criticalthinking.net/howteach.html

Ennis, R. (2018). Critical thinking across the curriculum: A vision. *Topoi*, 37(1), 165-184. DOI:10.1007/s11245-016-9401-4

Ennis, R., & Chattin, G. (2015). *An annotated list of English-language critical thinking tests.* Retrieved from www.criticalthinking.net/TestListDraft050315.docx

Ennis, R., & Weir, E. (1985). *The Ennis–Weir Critical Thinking Essay Test: An instrument for teaching and testing*. Pacific Grove, California: Midwest Publications.

Ennis, R., Millman, J., & Tomko, T. (1985). *Cornell Critical Thinking Tests Level X And Level Z Manual* (3rd ed.). Pacific Grove, California: Midwest Publications.

Erdogan, F. (2019). Effect of cooperative learning supported by reflective thinking activities on students' critical thinking skills. *Eurasian Journal of Educational Research*, 19(80), 89-112.

Erduran, S., & Kaya, E. (2019). Philosophy of chemistry and chemistry education. In S. Erduran & E. Kaya (Eds.), *Transforming teacher education through the epistemic core of chemistry* (pp. 1-24). Cham, Switzerland: Springer.

Erikson, M., & Erikson, M. (2018). Learning outcomes and critical thinking – Good intentions in conflict. *Studies in Higher Education,* 1-11. DOI:10.1080/03075079.2018.1486813

Espinosa, A., Monterola, S., & Punzalan, A. (2013). Career-oriented performance tasks in chemistry: Effects on students' critical thinking skills. *Education Research International*, Article 834584, 1-10. DOI:10.1155/2013/834584

Etzkowitz, H., Schuler, E., & Gulbrandsen, M. (2000). The evolution of the entrepreneurial university. In M. Jacob & T. Hellström (Eds.), *The future of knowledge production in the academy* (pp. 40-60). Buckingham: SRHE and Open University Press.

Evans, C., Kandiko Howson, C., & Forsythe, A. (2018). Making sense of learning gain in higher education. *Higher Education Pedagogies*, 3(1), 1-45. DOI:10.1080/23752696.2018.1508360

Evenhouse, D., Zadoks, A., Silva de Freitas, C., Patel, N., Kandakatla, R., Stites, N., Prebel, T., Berger, E., Krousgrill, C., Rhoads, J., & DeBoer, J. (2018). Video coding of classroom observations for research and instructional support in an innovative learning environment. *Australasian Journal of Engineering Education,* 23(2), 95-105. DOI:10.1080/22054952.2018.1519984

Ewert, A., & Sibthorp, J. (2000). Multivariate analysis in experiential education: Exploring the possibilities. *Journal of Experiential Education*, 23(2), 108-117. DOI:10.1177/105382590002300209

Facione, P. (1990a). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction—Executive summary of the Delphi report*. Millbrae: California Academic Press.

Facione, P. (1990b). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Research findings and recommendations*. Newark, Delaware: American Philosophical Association. Retrieved from https://philarchive.org/archive/FACCTA

Facione, P. (1990c). *The California Critical Thinking Skills Test.* Millbrae, California: California Academic Press.

Facione, P. (2000). The disposition toward critical thinking: Its character, measurement, and relation to critical thinking skill. *Informal Logic*, 20(1), 61-84. DOI:10.22329/il.v20i1.2254

Facione, P. (2015). Critical thinking: What it is and why it counts. *Insight assessments*. Retrieved from http://www.insightassessment.com/Resources/Select-Tools-For-Teaching-For-and-About-Thinking/Critical-Thinking-What-It-Is-and-Why-It-Counts/Critical-Thinking-What-It-Is-and-Why-It-Counts-PDF

Fan, W., & Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review. *Computers in human behavior*, 26(2), 132-139. DOI:10.1016/j.chb.2009.10.015

Fensham, P., & Bellocchi, A. (2013). Higher order thinking in chemistry curriculum and its assessment. *Thinking Skills and Creativity,* 10, 250-264. DOI:10.1016/j.tsc.2013.06.003

Fillion, N., & Martelli, D. (2017). *Overview of critical thinking in the BC K-12 Curriculum Revisions: Implications for post-secondary teaching and learning*. Technical report. Burnaby, Canada: Simon Fraser University. Retrieved from https://www.sfu.ca/conferences/criticalthinking/pastconf/devmindsfeb18/positionpaper.html

Finney, S., Myers, A., &Mathers, C. (2017). Test instructions do not moderate the indirect effect of perceived test importance on test performance in low-stakes testing contexts. *International Journal of Testing*, 18(4), 297-322. DOI:10.1080/15305058.2017.1396466

Finney, S., Sundre, D., Swain, M., & Williams, L. (2016). The validity of value-added estimates from low-stakes testing contexts: The impact of change in test-taking motivation and test consequences. *Educational Assessment*, 21(1), 60-87. DOI:10.1080/10627197.2015.1127753

Finocchairo, M. (1996). Critical Thinking, Critical Reflection and Methodological Reflection. *Inquiry: Critical Thinking Across the Disciplines*, 15(4), 66-79. DOI:10.5840/inquiryctnews19961546

Fisher, A. (2011). *Critical thinking: An introduction* (2nd ed.). Cambridge: Cambridge University Press.

Flavell, J. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10), 906-911. DOI:10.1037/0003-066X.34.10.906

Florea, N., & Hurjui, E. (2015). Critical thinking in elementary school children. *Procedia Social and Behavioral Sciences*, 180, 565-572. DOI:10.1016/j.sbspro.2015.02.161

Flores, K., Matking, G., Burbach, M., Quinn, C., & Harding, H. (2012). Deficient critical thinking skills among college graduates: Implications for leadership. *Educational Philosophy and Theory*, 44(2), 212-230. DOI:10.1111/j.1469-5812.2010.00672.x

Flynn, A. (2011). Developing problem-solving skills through retrosynthetic analysis and clickers in organic chemistry. *Journal of Chemical Education*, 88(11), 1496-1500. DOI:10.1021/ed200143k

Follmer, J., & Sperling, R. (2016). The mediating role of metacognition in the relationship between executive function and self-regulated learning. *British Journal of Educational Psychology,* 86(4), 559-575. DOI:10.1111/bjep.12123

Fong, C., Kim, Y., Davis, C., Hoang, T., & Kim, Y. (2017). A meta-analysis on critical thinking and community college student achievement. *Thinking Skills and Creativity,* 26, 71-83. DOI:10.1016/j.tsc.2017.06.002

Forawi, S. (2016). Standard-based science education and critical thinking. *Thinking Skills and Creativity,* 20, 52-62. DOI:10.1016/j.tsc.2016.02.005

Fortuin, K., & van Koppen, C. (2015). Teaching and learning reflexive skills in inter- and transdisciplinary research: A framework and its application in environmental science education. *Environmental Education Research*, 22(5), 697–716. DOI:10.1080/13504622.2015.1054264

Fortus, D. (2014). Attending to affect. *Journal of Research in Science Teaching*, 51(7), 821–835. DOI:10.1002/tea.21155

Fowler, F. (2009). *Survey research methods* (4th ed.). Thousand Oaks, California: Sage Publications. DOI:10.4135/9781452230184

Franco, A., & Almeida, L. (2015). Real-World Outcomes and critical thinking: differential analysis by academic major and gender. *Paidéia (Ribeirão Pret*o), 25(61), 173-181. DOI:10.1590/1982-43272561201505

Franco, A., Costa, P., Butler, H., & Almeida, L. (2017). Assessment of undergraduates' real-world outcomes of critical thinking in everyday situations. *Psychological reports*, 120(4), 707-720. DOI:10.1177/0033294117701906

Fredette, E. (2018). *Critical thinking skills as a predictor of student success in general chemistry: A correlational study* (Doctoral dissertation, Capella University, Minneapolis, Minnesota). Retrieved from ProQuest Dissertations & Theses Global. (UMI 10751866).

Freeman, S., Eddy, S., McDonough, M., Smith, M., Okoroafor, N., Jordt, H., & Wenderoth, M. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410-8415. DOI:10.1073/pnas.1319030111

Frezell, D. (2018). *Impact of inquiry based learning on students' motivation, engagement and attitude in science* (Masters dissertation, University of Windsor, Windsor, Ontario). Retrieved from ProQuest Dissertations & Theses Global. (UMI 10681683).

Friedlan, J. (1995). The effects of different teaching approaches on students' perceptions of the skills needed for success in accounting courses and by practicing accountants. *Issues in Accounting Education,* 10(1), 47- 63. Retrieved from https://search.proquest.com/docview/210921212?accountid=10910

Fukuzawa, S., & deBraga, M. (2019). Graded Response Method: Does question type influence the assessment of critical thinking? *Journal of Curriculum and Teaching*, 8(1), 1-10. DOI: doi.org/10.5430/jct.v8n1p1

Fung, D. (2017). The pedagogical impacts on students' development of critical thinking dispositions: Experience from Hong Kong secondary schools. *Thinking Skills and Creativity,* 26, 128-139. DOI:10.1016/j.tsc.2017.10.005

Fung, L., & Chow, L. (2002). Congruence of student teachers' pedagogical images and actual classroom practices. *Educational Research,* 44, 313–321. DOI:10.1080/00131880022000031605

Gabriel, F., Signolet, J., & Westwell, M. (2018). A machine learning approach to investigating the effects of mathematics dispositions on mathematical literacy. *International Journal of Research & Method in Education*. 41(3), 306-327. DOI:10.1080/1743727X.2017.1301916

Gajda, A., Karwowski, M., & Beghetto, R. (2017). Creativity and academic achievement: A meta-analysis. *Journal of Educational Psychology,* 109(2), 269. DOI:10.1037/edu0000133

Gallagher, C., Hipkins, R., & Zohar, A. (2012). Positioning thinking within national curriculum and assessment systems: Perspectives from Israel, New Zealand and Northern Ireland. *Thinking Skills and Creativity,* 7(2), 134-143. DOI:10.1016/j.tsc.2012.04.005

Gallagher, P. (2009). Collaborative essay testing: group work that counts. *International Journal of Nursing Education Scholarship*, 6(1), 1-11. DOI:10.2202/1548-923X.1845

Gambrill, E. (2005). *Critical thinking in clinical practice: A critical thinker's guide* (2nd ed.). New York: Wiley.

Ghanaat Pisheh, E., NejatyJahromy, Y., Gargari, R., Hashemi, T., & Fathi-Azar, E. (2019). Effectiveness of clicker-assisted teaching in improving the critical thinking of adolescent learners. *Journal of Computer Assisted Learning*, 35(1), 82-88. DOI:10.1111/jcal.12313

Gelerstein, D., del Río, R., Nussbaum, M., Chiuminatto, P., & López, X. (2016). Designing and implementing a test for measuring critical thinking in primary school. *Thinking Skills and Creativity,* 20, 40-49. DOI:10.1016/j.tsc.2016.02.002

Gellin, A. (2003). The effect of undergraduate student involvement on critical thinking. *Journal of College Student Development*, 44, 746-762. DOI:10.1353/csd.2003.0066

Ghanizadeh, A. (2017). The interplay between reflective thinking, critical thinking, self-monitoring, and academic achievement in higher education. *Higher Education*, 74, 101-114. DOI:10.1007/s10734-016-0031-y

Giancarlo, C., & Facione, P. (2001). A look across four years at the disposition toward critical thinking among undergraduate students. *The Journal of General Education,* 50(1), 29-55. DOI:10.1353/jge.2001.0004

Gibbs, P., Cartney, P., Wilkinson, K., Parkinson, J., Cunningham, S., James-Reynolds, C., Zoubir, T., Brown, V., Barter, P., Sumner, P., & MacDonald, A. (2017). Literature review on the use of action research in higher education. *Educational Action Research,* 25(1), 3-22. DOI:10.1080/09650792.2015.1124046

Gilbert, J. (2015). *Science education.* In J. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (2nd ed., vol. 2, pp. 229-235). Amsterdam: Elsevier. DOI:10.1016/B978-0-08-097086-8.92098-1

Gilboy, N., & Kane, D. (2004). Unfolding case based scenarios: A method of teaching and testing the critical thinking skills of newly licensed nurses. *Journal of Emergency Nursing*, 30(1), 83–85. DOI:10.1016/j.jen.2003.11.007

Glaser, E. (1941). An experiment in the development of critical thinking. *Teachers College Record*, 43(5), 409–410.

Glaser, R. (1983). *Education and thinking: The role of knowledge*. Technical Report No. PDS-6, 1-39. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a130532.pdf

Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist*, 39, 93–104. DOI:10.1037/0003-066X.39.2.93

Glynn, S., Taasoobshirazi, G., & Brickman, P. (2007). Nonscience majors learning science: A theoretical model of motivation. *Journal of Research in Science Teaching,* 44(8), 1088–1107. DOI:10.1002/tea.20181

Gojkov, G., Stojanović, A., & Rajić, A. (2015). Critical thinking of students–indicator of quality in higher education. *Procedia Social and Behavioral Sciences*, 191, 591-596. DOI:10.1016/j.sbspro.2015.04.501

Golding, C. (2011). Educating for critical thinking: thought- encouraging questions in a community of inquiry, *Higher Education Research and Development,* 30(3), 357-370. DOI:10.1080/07294360.2010.499144

Goodin, H. (2005). The *use of deliberative discussion as a teaching strategy to enhance the critical thinking abilities of freshman nursing students* (Doctoral dissertation, Ohio University, Columbus, Ohio). Retrieved from ProQuest Dissertations & Theses Global. (UMI 3176411).

Gooding, J. (2009). *Comparing the perceptions of scientific inquiry between experts and practitioners* (Doctoral dissertation, Robert Morris University, Moon, Pennsylvania). Retrieved from ProQuest Dissertations & Theses Global. (UMI 3383001).

Gordon, E., & Campbell, E. (2014). Context and perspective: Implications for assessment in education. *Teachers College Record*, 116(11), 1-18. Retrieved from https://www.ets.org/Media/Research/pdf/gordon_campbell_implications_assessment_education.pdf

Gorman, B. (2012). *Governmentality in higher education: A critical analysis of the National Survey of Student Engagement (NSSE)* (Doctoral dissertation, Michigan Technological University, Houghton, Michigan). Retrieved from ProQuest Dissertations & Theses Global. (UMI 3509744).

Gottesman, A., & Hoskins, S. (2013). CREATE Cornerstone: Introduction to scientific thinking, a new course for STEM-interested freshmen, demystifies scientific thinking through analysis of scientific literature. *CBE-Life Sciences Education,* 12(1), 59-72. DOI:10.1187/cbe.12-11-0201

Gouvea, J. (2017). Insights from small-N studies. *CBE-Life Sciences Education*, 16(3), 1-3. DOI:10.1187/cbe.17-06-0110

Government of South Australia (2015). *SACE Capabilities Policy*. South Australian Certificate of Education, SACE Board of SA. Retrieved from: https://www.sace.sa.edu.au/students/sace-overview/sace-capabilities

Grant, M., & Smith, M. (2018). Quantifying Assessment of Undergraduate Critical Thinking. *Journal of College Teaching & Learning*, 15(1), 27-38. DOI:10.19030/tlc.v15i1.10199

Greene, J., & Yu, S. (2016). Educating Critical Thinkers: The Role of Epistemic Cognition. *Policy Insights from Behavioural and Brain Sciences*, 3(1), 45-53. DOI:10.1177/2372732215622223

Greiff, S., Niepel, C., & Wistenberg, S. (2015). 21st century skills: International advancements and recent developments. *Thinking Skills and Creativity,* 18, 1–3. DOI:10.1016/j.tsc.2015.04.007

Greiff, S., Wüstenberg, S., Csapó, B., Demetriou, A., Hautamäki, J., Graesser, A., & Martin, R. (2014). Domain-general problem solving skills and education in the twenty-first century. *Educational Research Review*, 13, 74-83. DOI:10.1016/j.edurev.2014.10.002

Grieco, L. (2016). *Meeting the Demands of the twenty-first Century Workplace: Effects of Critical Thinking Instruction on the Application of Critical Thought* (Doctoral dissertation, North Central University, Minneapolis, Minnesota). Retrieved from ProQuest Dissertations & Theses Global. (UMI 10103879).

Gross, M. (2017). The dangers of a post-truth world. *Current Biology*, 27(1), R1–R4. DOI:10.1016/j.cub.2016.12.034

Groves, R., Cialdini, R., & Couper, M. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly*, 56(4), 475-495. DOI:10.1086/269338

Groves, R., Presser, S., & Dipko, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly*, 68(1), 2-31. DOI:10.1093/poq/nfh002

Grussendorf, J., & Rogol, N. (2018). Reflections on critical thinking: Lessons from a quasi-experimental study. *Journal of Political Science Education*, 14(2), 151-166. DOI:10.1080/15512169.2017.1381613

Gupta, T., Burke, K., Mehta, A., & Greenbowe, T. (2015). Impact of guided-inquiry-based instruction with a writing and reflection emphasis on chemistry students' critical thinking abilities. *Journal of Chemical Education*, 92, 32–38. DOI:10.1021/ed500059r

Haak, D., HilleRisLambers, J., Pitre, E., & Freeman, S. (2011). Increased structure and active learning reduce the achievement gap in introductory biology. *Science*, 332(6034), 1213-1216. DOI:10.1126/science.1204820

Hadorn, G., Biber-Klemm, S., Grossenbacher-Mansuy, W., Hoffmann-Riem, H., Joye, D., Pohl, C Wiesmann, U., & Zemp, E. (Eds.). (2008). *Handbook of transdisciplinary research*. Dordrecht, Netherlands: Springer.

Haller, E., Child, D., & Walberg, H. (1988). Can comprehension be taught? A quantitative synthesis of metacognitive studies. *Educational Researcher,* 17(9), 5-8. DOI:10.3102/0013189X017009005

Halpern, D. (1998). Teaching critical thinking for transfer across domains: Dispositions, skills, structure training, and metacognitive monitoring. *American Psychologist*, 53(4), 449-455. DOI:10.1037/0003-066X.53.4.449

Halpern, D. (1999). Teaching for critical thinking: Helping college students develop the skills and dispositions of a critical thinker. *New Direction for Teaching and Learning*. 80, 69-74. DOI:10.1002/tl.8005

Halpern, D. (2001). Assessing the effectiveness of critical thinking instruction. *The Journal of General Education,* 50(4), 270–286. Retrieved from https://www.jstor.org/stable/27797889

Halpern, D. (2003). *Knowledge and thought: An introduction to critical thinking.* Hillsdale, New Jersey: Lawrence Erlbaum.

Halpern, D. (2010). *Halpern Critical Thinking Assessment*. Schuhfried (Vienna Test System). http://www.schuhfried.at.

Halpern, D. (2013) A is for assessment: The other scarlet letter. *Teaching of Psychology*, 40(4), 358-362. DOI:10.1177/0098628313501050

Halpern, D. (2014). *Thought and knowledge* (5th ed.). New York: Psychology Press.

Halpern, D., Millis, K., Graesser, A., Butler, H., Forsyth, C., & Cai, Z. (2012). Operation ARA: A computerized learning game that teaches critical thinking and scientific reasoning. *Thinking Skills and Creativity,* 7(2), 93-100. DOI:10.1016/j.tsc.2012.03.006

Harper, D. (2018a). 'critic (n.)'. Online Etymology Dictionary. Retrieved 21 March 2018, <https://www.etymonline.com/word/critic>

Harper, D. (2018b). '*krei'. Online Etymology Dictionary. Retrieved 21 March 2018, https://www.etymonline.com/word/*krei-

Harris, K. (2015). *Assessing course impacts on critical thinking: The relationship between self-reported learning gains and performance* (Doctoral dissertation, Tennessee Technological University, Cookeville, Tennessee). Retrieved from ProQuest Dissertations & Theses Global. (UMI 3705421).

Harris, K., Stein, B., Haynes, A., Lisic, E, & Leming, K. (2014). Identifying courses that improve students' critical thinking skills using the CAT Instrument: A Case Study. *Proceedings of the 10th Annual International Joint Conference on Computer, Information, System Sciences and Engineering*, 10, 1-4. Retrieved from https://www.tntech.edu/cat/pdf/reports/Harris_EIAE.pdf

Hatcher, D. (2000). Arguments for another definition of critical thinking. *Inquiry: Critical Thinking Across the Disciplines*, 20(1), 3-8. DOI:10.5840/inquiryctnews20002016

Hatcher, D. (2009). The institutional assessment of critical thinking: a fifteen-year perspective. In J. Sobocan & L. Groarke (Eds.), *Critical thinking education and assessment: can high order thinking be tested?* (pp. 229-246). Aylmer, Ontario: Althouse Press.

Hatcher, D. (2013). Reflections on critical thinking: theory, practice and assessment. *Inquiry: Critical Thinking Across the Disciplines*, 28(2), 4-24. DOI:10.5840/inquiryct20132827

Hattie, J. (1992). Measuring the effects of schooling. *Australian Journal of Education,* 36, 5-13. DOI:10.1177/000494419203600102

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. DOI:10.3102/003465430298487

Haught-Tromp, C. (2017). The green eggs and ham hypothesis: How constraints facilitate creativity. *Psychology of Aesthetics, Creativity, and the Arts,* 11(1), 10-17. DOI:10.1037/aca0000061

Hawkins, K. (2012). *Thinking and Reading Among College Undergraduates: An Examination of the Relationship between Critical Thinking Skills and Voluntary Reading* (Doctoral dissertation, University of Tennessee, Knoxville, Tennessee). Retrieved from http://trace.tennessee.edu/utk_graddiss/1302

Hays, S. (2019). Nature as discourse: Transdisciplinarity and vagus nerve function. *Transdisciplinary Journal of Engineering & Science*, 10, 18-27. DOI:10.22545/2019/0112

Head, A. (2013, April). Project Information Literacy: What can be learned about the information-seeking behavior of today's college students? In D. Mueller (Ed.), *Imagine, innovate, inspire: The Proceedings of the ACRL 2013 Conference* (pp. 472-48). Indianapolis, Indiana: Association of College and Research Libraries.

Heijltjes, A., Van Gog, T., & Paas, F. (2014). Improving students' critical thinking: Empirical support for explicit instructions combined with practice. *Applied Cognitive Psychology*, 28(4), 518-530. DOI:10.1002/acp.3025

Heilig, J., Brewer, T., & Pedraza, J. (2018). Examining the myth of accountability, high-stakes testing, and the achievement gap. *Journal of Family Strengths*, 18(1), Article 9, 1-14. https://digitalcommons.library.tmc.edu/jfs/vol18/iss1/9

Helbing, D., Frey, B., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., Van Den Hoven, J., Zicari, R., & Zwitter, A. (2019). Will democracy survive big data and artificial intelligence? In D. Helbing (Ed.), *Towards Digital Enlightenment* (pp. 73-98). Cham, Switzerland: Springer. DOI:10.1007/978-3-319-9086-4_7

Henderson, C., Beach, A., & Finkelstein, N. (2011). Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature. *Journal of Research in Science Teaching,* 48(8), 952-984. DOI:10.1002/tea.20439

Henderson, C., & Fisher, K. (2018). Change in Higher Education. In C. Henderson, C. Rasmussen, A. Knaub, N. Apkarian, A. Daly & K. Fisher (Eds.), *Researching and enacting change in postsecondary education: Leveraging instructors' social networks* (Vol. 28, pp. 17-31). New York: Routledge.

Henderson, C., Connolly, M., Dolan, E. L., Finkelstein, N., Franklin, S., Malcom, S., Rasmussen, C., Redd, K., & St. John, K. (2017). Towards the STEM DBER Alliance: Why we need a discipline-based STEM education research community. *International Journal of Research in Undergraduate Mathematics Education*, 3(2), 247–254. DOI:10.1007/s40753-017-0056-3

Henderson, D. (2010). A chemical instrumentation game for teaching critical thinking and information literacy in instrumental analysis courses. *Journal of Chemical Education*, 87(4), 412-415. DOI:10.1021/ed800110f

Heng, C., & Ziguang, Y. (2015). Framework of assessment for the evaluation of thinking skills of tertiary level students. *Advances in Language and Literary Studies*, 6(5), 67-72. DOI:10.7575/aiac.alls.v.6n.5p.67

Hennessey, M. (1999, March). *Probing the dimensions of metacognition: Implications for conceptual change teaching-learning.* Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, Boston, Massachusetts.

Herman, G., & Mena, I. (2015, October). Tracking the spread of research-based instructional strategies. *IEEE Frontiers in Education Conference*, El Paso, Texas. DOI:10.1109/fie.2015.7344218

Higgins, S. (2016). *Managing Academic Libraries: Principles and Practice*. Cambridge, Massachusetts: Elsevier. DOI:10.1016/B978-1-84334-621-0.00004-2

Higgins, S., Kokotsaki, D., & Coe, R. (2012). The *Teaching and Learning Toolkit: Technical Appendices*. London: The Education Endowment Foundation and Sutton Trust. Retrieved from http://educationendowmentfoundation.org.uk/uploads/pdf/Technical_Appendices_(July_2012).pdf

Hirsh, S. (1997). How do children find information on different types of tasks? Children's use of the science library catalog, *Library Trends,* 45(4), 725–745. Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/8125/librarytrendsv45i4l_opt.pdf

Hirsh, S. (2004). Domain knowledge and children's search behaviour. In M. Chelton and C. Cool (Eds.), *Youth information-seeking behavior: Theories, models, and issues* (Vol. 1, pp.241-270). Lanham, Maryland: Scarecrow Press, Inc.

Hirsh, S. (2018). The transformative information landscape. In S. Hirsh (Ed.), *Information services today: An introduction* (2nd ed., pp. 3-13). Lanham, Maryland: Rowman & Littlefield.

Hitchcock, D. (2015). The effectiveness of instruction in critical thinking. In M. Davies & R. Barnett (Eds.), *The Palgrave handbook of critical thinking in higher education* (pp. 283-294). New York: Palgrave Macmillan.

Hoffmann-Riem, H., Biber-Klemm, S., Grossenbacher-Mansuy, W., Hadorn, G., Joye, D., Pohl, C., Wiesmann, U., & Zemp, E. (2008). Idea of the handbook. In G. Hadorn, S. Biber-Klemm, W. Grossenbacher-Mansuy, H. Hoffmann-Riem, D. Joye, C. Pohl, U. Wiesmann & E. Zemp, (Eds.), *Handbook of transdisciplinary research* (pp. 3-17). Dordrecht, Netherlands: Springer.

Holmes, N., Wieman, C., & Bonn, D. (2015). Teaching critical thinking. *Proceedings of the National Academy of Sciences*, 112(36), 11199-11204. DOI:10.1073/pnas.1505329112

Holt, E., Young, C., Keetch, J., Larsen, S., & Mollner, B. (2015). The greatest learning return on your pedagogical investment: Alignment, assessment or in-class instruction? *PLOS One*, 10(9), e0137446. DOI:10.1371/journal.pone.0137446

Hopfenbeck, T. (2018). Classroom assessment, pedagogy and learning – twenty years after Black and Wiliam 1998. *Assessment in Education: Principles, Policy & Practice*, 25(6), 545–550. DOI:10.1080/0969594x.2018.1553695

Hora, M. (2014). Limitations in experimental design mean that the jury is still out on lecturing. *Proceedings of the National Academy of Sciences*, 111(30), E3024. DOI:10.1073/pnas.1410115111

Hubbard, J. (2017). *Postsecondary instructor attitudes toward tablet use for collaboration and critical thinking development* (Doctoral dissertation, Walden University, Minneapolis, Minnesota). Retrieved from ProQuest Dissertations & Theses Global. (UMI 10634340).

Huber, C., & Kuncel, N. (2016). Does college teach critical thinking? A meta-analysis. *Review of Educational Research*, 86(2), 431-468. DOI:10.3102/0034654315605917

Hugerat, M., & Kortam, N. (2014). Improving higher order thinking skills among freshmen by teaching science through inquiry. *Eurasia Journal of Mathematics, Science and Technology Education*, 10, 447–454. DOI:10.12973/eurasia.2014.1107a

Hunter, N. (2014). *General topic guidelines: BIOL1112 Biology and Society*. Flinders University, Bedford Park, South Australia.

Hunter, N. (2015). *Statement of Assessment Methods. BIOL1112 Biology and Society*. Flinders University, Bedford Park, South Australia.

Hurd, P. (2001). Modernizing science education. *Journal of Research in Science Teaching,* 39(1), 3–9. DOI:10.1002/tea.10002

Husain, H., Bais, B., Hussain, A., & Samad, S. (2012). How to construct open ended questions. *Procedia-Social and Behavioral Sciences*, 60, 456-462. DOI:10.1016/j.sbspro.2012.09.406

Hyytinen, H., Nissinen, K., Ursin, J., Toom, A., & Lindblom-Ylänne, S. (2015). Problematising the equivalence of the test results of performance-based critical thinking tests for undergraduate students. *Studies in Educational Evaluation*, 44, 1-8. DOI:10.1016/j.stueduc.2014.11.001

Hyytinen, H., Toom, A., & Postareff, L. (2018). Unraveling the complex relationship in critical thinking, approaches to learning and self-efficacy beliefs among first-year educational science students. *Learning and Individual Differences,* 67, 132-142. DOI:10.1016/j.lindif.2018.08.004

İlhan, M., & Gezer, M. (2017). A comparison of the reliability of the Solo-and revised Bloom's Taxonomy-based classifications in the analysis of the cognitive levels of assessment questions. *Pegem Eğitim ve Öğretim Dergisi*, 7(4), 637-662. DOI:10.14527/pegegog.2017.023

Inouye, J., & Flannelly, L. (1998). Inquiry-based learning as a teaching strategy for critical thinking. *Clinical Nurse Specialist*, 12(2), 67-72. DOI:10.1097/00002800-199803000-00009

Insight Assessment (2018). *California Critical Thinking Skills Test (CCTST).* San Jose, California: The California Academic Press LLC. Retrieved from https://www.insightassessment.com/Products/Products-Summary/Critical-Thinking-Skills-Tests/California-Critical-Thinking-Skills-Test-CCTST

Ironside, P. (2004). "Covering content" and teaching thinking: Deconstructing the additive curriculum. *Journal of Nursing Education*, 43(1), 5-12. DOI:10.3928/01484834-20040101-02

Jacob, C. (2004). Critical thinking in the chemistry classroom and beyond. *Journal of Chemical Education*, 81(8), 1216-1223. DOI:10.1021/ed081p1216

Jang, H. (2015). Identifying 21st century STEM competencies using workplace data. *Journal of Science Education and Technology*, 25(2), 284–301. DOI:10.1007/s10956-015-9593-1

Jensen, J., Kummer, T., & Godoy, P. (2015). Improvements from a flipped classroom may simply be the fruits of active learning. *CBE—Life Sciences Education*, 14(1), ar5, 1-12. DOI:10.1187/cbe.14-08-0129

Johnson, R. (1996). Critical thinking and command of language. *Inquiry: Critical Thinking Across the Disciplines*, 16(2), 78-92. DOI:10.5840/inquiryctnews199616215

Johnson, R., & Hamby, B. (2015). A meta-level approach to the problem of defining "critical thinking." *Argumentation*, 29(4), 417–430. DOI:10.1007/s10503-015-9356-4

Jones, A. (2006). *Re-disciplining generic skills: An examination of the relationship between the disciplinary context and generic skills in higher education* (Doctoral dissertation, University of Melbourne, Parkville, Victoria). Retrieved from http://cat.lib.unimelb.edu.au/record=b3130165

Jones, A. (2007). Multiplicities or manna from heaven? Critical thinking and the disciplinary context. *Australian Journal of Education*, 51(1), 84-103. DOI:10.1177/000494410705100107

Jones, A. (2009). Re-disciplining generic attributes: The disciplinary context in focus. *Studies in Higher Education,* 34(1), 85-100. DOI:10.1080/03075070802602018

Jones, A. (2013). There is nothing generic about graduate attributes: unpacking the scope of context. *Journal of Further and Higher Education*, 37(5), 591-605. DOI:10.1080/03075070802602018

Jones, S., Yates, B., & Kelder, J. (2011). *Science learning and teaching academic standards statement.* Strawberry Hills, New South Wales: Australian Learning and Teaching Council. Retrieved from http://www.olt.gov.au/resources

Jungwirth, E. (1987). Avoidance of Logical Fallacies: a neglected aspect of science-education and science-teacher education. *Research in Science & Technological Education*, 5(1), 43-58. DOI:10.1080/0263514870050106

Jungwirth, E., & Dreyfus, A. (1990). Diagnosing the attainment of basic enquiry skills: The 100-year old quest for critical thinking. *Journal of Biological Education*, 24(1), 42-49. DOI:10.1080/00219266.1990.9655100

Kadian, A. (2016). Relation between academic achievement, executive function, intelligence and meta cognition. *International Journal of Indian Psychology*, 3(4), 166-174. DOI:10.25215/0304.074

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kaptan, K., & Timurlenk, O. (2012). Challenges for science education. *Procedia Social and Behavioral Sciences*, 51, 763-771. DOI:10.1016/j.sbspro.2012.08.237

Karabulut, U. (2015). Identifying instructional methods of teaching critical thinking: A systematic review and analysis of three decades of literature. *The Georgia Social Studies Journal*, 5(2), 96-107. Retrieved from https://coe.uga.edu/assets/downloads/misc/gssj/Karabulut-2015.pdf

Kennedy, M., Fisher, M., & Ennis, R. (1991). Critical thinking: Literature review and needed research. In L. Idol & B. Jones (Eds.), *Educational values and cognitive instruction: Implications for reform* (pp. 11-40). Hillsdale, New Jersey: Lawrence Erlbaum and Associates.

Kezar, A., Holcombe, E., & Kitchen, J. (2018). *Scaling Change in Higher Education: A Guide for External Stakeholder Groups*. Los Angeles, California: Pullias Center for Higher Education.

Kim, K., Sharma, P., Land, S., & Furlong, K. (2013). Effects of Active Learning on Enhancing Student Critical Thinking in an Undergraduate General Science Course. *Innovative Higher Education,* 38(3), 223–235. DOI:10.1007/s10755-012-9236-x

Kinchin, I. (2019). Pedagogic Frailty and the Ecology of Teaching at University: A Case of Conceptual Exaptation. In A. Railean (Ed.), *Handbook of Research on Ecosystem-Based Theoretical Models of Learning and Communication* (pp. 154-166). Hershey, Pennsylvania: IGI Global.

King, F., Goodson, L., & Rohani, F. (1998). *Higher Order Thinking Skills: Definitions, strategies, assessment.* Tallahasee, Florida: Center for the Advancement of Learning and Assessment Florida State University.

King, P., & Kitchener, K. (1994). *Developing Reflective Judgment: Understanding and Promoting Intellectual Growth and Critical Thinking in Adolescents and Adults*. San Francisco, California: Jossey-Bass Publishers.

Kinsbourne, M. (2001). Mind and Nature: Essays on Time and Subjectivity, *Neuropsychoanalysis*, 3(2), 253-255. DOI:10.1080/15294145.2001.10773363

Kinsbourne, M. (2004). Commentary on "The Pleasantness of False Beliefs", *Neuropsychoanalysis*, 6(1), 33-37. DOI:10.1080/15294145.2004.10773437

Kinslow, A., & Sadler, T. (2018). Making Science Relevant: Using Socio-Scientific Issues to Foster Critical Thinking. *The Science Teacher*, 86(1), 40-45. Retrieved from https://www.questia.com/read/1G1-549158301/making-science-relevant-using-socio-scientific-issues

Kirkpatrick, D. (1996). *Great ideas revisited: Techniques for evaluating training programs*. Madison, Wisconsin: American Society for Training and Development.

Kirmizi, F., Saygi, C., & Yurdakal, I. (2015). Determine the relationship between the disposition of critical thinking and the perception about problem solving skills. *Procedia Social and Behavioral Sciences*, 191, 657-661. DOI:10.1016/j.sbspro.2015.04.719

Kirpatrick, D (1975*). Evaluating training programs*. Madison, Wisconsin: American Society for Training and Development.

Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes*, Cambridge, Massachusetts: MIT Press.

Klein, J. (2013). The transdisciplinary moment (um). *Integral Review*, 9(2). Retrieved from http://www.integral-review.org/issues/vol_9_no_2_jun_2013_full_issue.pdf#page=194

Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The Collegiate Learning Assessment: Facts and fantasies. *Evaluation Review*, 31(5), 415-439. DOI:10.1177/0193841X07303318

Kobzeva, N. (2015). Scrabble as a tool for engineering students' critical thinking skills development. *Procedia Social and Behavioral Sciences*, 182, 369-374. DOI:10.1016/j.sbspro.2015.04.791

Koeper, I. (2015). *Statement of Assessment Methods. CHEM1101 Chemical Structure and Bonding***.** Flinders University, Bedford Park, South Australia.

Kong, S. (2006). Using Thinking Skills to Intrinsically Motivate Effective Learning. *The International Journal of Creativity & Problem Solving*, 16(1), 75-89 Retrieved from http://www.dbpia.co.kr/Article/NODE01045511

Kong, S. (2015). An experience of a three-year study on the development of critical thinking skills in flipped secondary classrooms with pedagogical and technological support. *Computers & Education*, 89, 16-31. DOI:10.1016/j.compedu.2015.08.017

Korhonen, V., Inkinen, M., Mattsson, M., & Toom, A. (2017). Student engagement and the transition from the first to second year in higher education. In E. Kyndt, V. Donche, K. Trigwell & S. Lindblom-Ylänne (Eds.), *Higher education transitions: theory and research* (pp. 113-135). New York: Routledge.

Kovic, M. (2016). A generalized definition of critical thinking. *Swiss Skeptics Discussion Paper Series*, 1(1), 1-26. Retrieved from https://www.skeptiker.ch/wp-content/uploads/2016/12/Discussion-Paper-Critical-Thinking.pdf

Kowalczyk, N., Hackworth, R., & Case-Smith, J. (2012). Perceptions of the use of critical thinking teaching methods. *Radiologic Technology*, 83(3), 226-236. Retrieved from http://www.radiologictechnology.org/content/83/3/226.abstract

Kramarski, B., & Mevarech, Z. (2003). Enhancing mathematical reasoning in the classroom: The effects of cooperative learning and metacognitive training. *American Educational Research Journal*, 40(1), 281-310. DOI:10.3102/00028312040001281

Kranzfelder, P., Lo, A., Melloy, M., Walker, L., & Warfa, A. (2019). Instructional practices in reformed undergraduate STEM learning environments: a study of instructor and student behaviors in biology courses. *International Journal of Science Education*, 41(14), 1944-1961. DOI:10.1080/09500693.2019.1649503

Krapp, A. (1999). Interest, motivation and learning: An educational–psychological perspective. *European Journal of Psychology of Education*, 14, 23–40. DOI:10.1007/BF03173109

Krumm, G., Filippetti, V., & Gutierrez, M. (2018). The contribution of executive functions to creativity in children: What is the role of crystallized and fluid intelligence? *Thinking Skills and Creativity,* 29, 185-195. DOI:10.1016/j.tsc.2018.07.006

Ku, K. (2009). Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Thinking Skills and Creativity,* 4(1), 70-76. DOI:10.1016/j.tsc.2009.02.001

Ku, K., & Ho, I. (2010). Metacognitive strategies that enhance critical thinking. *Metacognition Learning*, 5(3), 251-267. DOI:10.1007/s11409-010-9060-6

Kuczynski, J., Costello, E., Nemergut, D., Zaneveld, J., Lauber, C., Knights, D., Koren, O. Fierer, N., Kelley, S., Ley, R., Gordon, J., & Knight, R. (2010). Direct sequencing of the human microbiome readily reveals community differences. *Genome biology*, 11(5), 210. DOI:10.1186/gb-2010-11-5-210

Kuh, G., Kinzie, J., Cruce, T., Shoup, R., & Gonyea, R. (2006). *Connecting the dots: Multi-faceted analyses of the relationships between student engagement results from the NSSE, and the institutional practices and conditions that foster student success.* Bloomington, Indiana: Indiana University, Center for Postsecondary Research. Retrieved from http://nsse.indiana.edu/pdf/Connecting_the_Dots_Report.pdf

Kuhn, D. (1999). A developmental model of critical thinking. *Educational Researcher,* 28(2), 16-26. DOI:10.3102/0013189X028002016

Kuhn, D. (2000). Metacognitive development. *Current Directions in Psychological Science,* 9(5), 178-181. DOI:10.1111/1467-8721.00088

Kuhn, D. (2019). Critical thinking as discourse. *Human Development*, 62(3), 146-164. DOI:10.1159/000500171

Kuhn, D., & Dean, D. (2004). Metacognition: A bridge between cognitive psychology and educational practice. *Theory into Practice,* 43(4), 268-273. DOI:10.1207/s15430421tip4304_4

Kuhn, D., Ramsey, S., & Arvidsson, T. (2015). Developing multivariable thinkers. *Cognitive Development,* 35, 92-110. DOI:10.1016/j.cogdev.2014.11.003

Kules, B. (2016). Computational thinking is critical thinking: Connecting to university discourse, goals, and learning outcomes. *Proceedings of the Association for Information Science and Technology*, 53(1), 1–6. DOI:10.1002/pra2.2016.14505301092

Kuncel, N. (2011). Measurement and meaning of critical thinking. *Research Report for the NRC twenty-first Century Skills Workshop*. Washington, DC: National Research Council.

Lai, E. (2011a). Critical thinking: A literature review. *Pearson's Research Reports,* 6. Retrieved from https://images.pearsonassessments.com/images/tmrs/CriticalThinkingReviewFINAL.pdf

Lai, E. (2011b). Metacognition: A literature review. *Pearson's Research Reports,* 24. Retrieved from https://images.pearsonassessments.com/images/tmrs/Metacognition_Literature_Review_Final.pdf

Lakomý, M., Hlavová, R., & Machackova, H. (2019). Open science and the science-society relationship. *Society,* 56(3), 246–255. DOI:10.1007/s12115-019-00361-w

Larsson, K. (2017). Understanding and teaching critical thinking—A new approach. *International Journal of Educational Research,* 84, 32-42. DOI:10.1016/j.ijer.2017.05.004

Laudonia, I., Mamlok-Naaman, R., Abels, S., & Eilks, I. (2017). Action research in science education – an analytical review of the literature. *Educational Action Research,* 26(3), 480–495. DOI:10.1080/09650792.2017.1358198

Lauer, T. (2005). Teaching critical-thinking skills using course content material: A reversal of roles. *Journal of College Science Teaching,* 34(6), 34-44. Retrieved from https://www.jstor.org/stable/42992377

Lautensach, A. (2004, December). *A tertiary curriculum for sustainability.* Paper presented at AARE Annual Conference, Melbourne, Australia. Retrieved from https://www.aare.edu.au/data/publications/2004/lau04260.pdf

Lautensach, A. (2017, November 27). An effective education for the future. *Policy Options.* Retrieved from https://policyoptions.irpp.org/magazines/november-2017/an-effective-education-for-the-future/

Lautensach, A., & Lautensach, S. (2010). A curriculum for a secure future: Agenda for reform. In G. Martin, D. Houston, P. McLaren & J. Suoranta (Eds.), *The havoc of capitalism: Publics, pedagogies and environmental crisis* (pp. 207-228). Rotterdam: Sense Publishers.

Leavy, P. (2016). *Essentials of transdisciplinary research: Using problem-centered methodologies*. New York: Routledge.

Lederman, L. (2008). Scientists and 21st century science education. *Technology in Society*, 30(3-4), 397-400. DOI:10.1016/j.techsoc.2008.04.014

Lee, M., Kim, H., & Kim, M. (2014). The effects of Socratic questioning on critical thinking in web-based collaborative learning. *Education as Change*, 18(2), 285-302. DOI:10.1080/16823206.2013.849576

Lee, Y. (2018). Nurturing critical thinking for implementation beyond the classroom: Implications from social psychological theories of behavior change. *Thinking Skills and Creativity,* 27, 139-146. DOI:10.1016/j.tsc.2018.02.003

Lehamn, A., O'Rourke, N., & Stepanski, L. (2005). *JMP for basic univariate and multivariate statistics*: *A step-by-step guide*. Cary, North Carolina: SAS Institute.

Leite, L., & Dourado, L. (2013). Laboratory activities, science education and problem-solving skills. *Procedia Social and Behavioral Sciences*, 106, 1677-1686. DOI:10.1016/j.sbspro.2013.12.190

Leming, K. (2016). *Effective instruction and assessment methods that lead to gains in critical thinking as measured by the Critical thinking Assessment Test (CAT)* (Doctoral dissertation, Tennessee Technological University, Cookeville, Tennessee). Retrieved from ProQuest Dissertations & Theses Global. (UMI 10112435).

Lennon, N. (2014). *How professors infuse critical thinking into college courses* (Doctoral dissertation, Seton Hall University, South Orange, New Jersey). Retrieved from ProQuest Dissertations & Theses Global. (UMI 3645747).

Lerer, N., & Talley, K. (2010). National Survey of Student Engagement's (NSSE) benchmarks–one size fits all? *On the Horizon*, 18(4), 355-363. DOI:10.1108/10748121011082653

Lewine, R., Sommers, A., Waford, R., & Robertson, C. (2015). Setting the Mood for Critical Thinking in the Classroom. *International Journal for the Scholarship of Teaching and Learning*, 9(2), Article 5. DOI:10.20429/ijsotl.2015.090205

Lew, M., Alwis, W., & Schmidt, H. (2010). Accuracy of students' self-assessment and their beliefs about its utility. *Assessment & Evaluation in Higher Education*, 35(2), 135-156. DOI:10.1080/02602930802687737

Lin, T., Lin, T., Potvin, P., & Tsai, C. (2018). Research trends in science education from 2013 to 2017: a systematic content analysis of publications in selected journals. *International Journal of Science Education*, 1–21. DOI:10.1080/09500693.2018.1550274

Lingard, B., Ladwig, J. Mills, M., Bahr, M., Chant, D., Warry, M., Aliwood, J., Capeness, R., Christie, P. Gore, J., Hayes, D., & Luke, A. (2001). *The Queensland School Reform Longitudinal Study*. Brisbane: Education Queensland.

Lipman, M. (1987). Critical thinking: What can it be? *Analytic Teaching*, 8(1), 5-12. Retrieved from http://journal.viterbo.edu/index.php/at/article/view/403

Lipman, M. (1995). Good thinking. *Inquiry: Critical Thinking Across the Disciplines*, 15(2), 37-41. DOI:10.5840/inquiryctnews199515224

Lipman, M. (2003). Thinking in education (2nd ed.). Cambridge: Cambridge University Press.

Lisic, E. (2015). *Creating change: Implementing the Critical thinking Assessment Test (CAT) as faculty development to improve instruction* (Doctoral dissertation, Tennessee Technological University, Cookeville, Tennessee). Retrieved from ProQuest Dissertations & Theses Global, (UMI 3722968).

Liu, O., Frankel, L., & Roohr, K. (2014). Assessing critical thinking in higher education: Current state and directions for next generation assessment. *ETS Research Report Series* (RR-14-10), 1, 1-23. DOI:10.1002/ets2.12009

Liu, O., Rios, J., & Borden, V. (2015). The effects of motivational instruction on college students' performance on low-stakes assessment. *Educational Assessment*, 20, 79–94. DOI:10.1080/10627197.2015.1028618

Llopart, M., & Esteban-Guitart, M. (2017). Strategies and resources for contextualising the curriculum based on the funds of knowledge approach: A literature review. *The Australian Educational Researcher*, 44(3), 255–274. DOI:10.1007/s1338 4-017-0237-8

Lombardi, A., Kowitt, J., & Staples, F. (2015). Correlates of critical thinking and college and career readiness for students with and without disabilities. *Career Development and Transition for Exceptional Individuals*, 38(3), 142-151. DOI:10.1177/2165143414534888

Lorencová, H., Jarošová, E., Avgitidou, S., & Dimitriadou, C. (2019). Critical thinking practices in teacher education programmes: a systematic review. *Studies in Higher Education,* 44(5), 844-859. DOI:10.1080/03075079.2019.1586331

Lovelace, M., & Brickman, P. (2013). Best practices for measuring students' attitudes toward learning science. *CBE-Life Sciences Education*, 12(4), 606–617. DOI:10.1187/cbe.12-11-0197

Lumpkin, A., Achen, R., & Dodd, R. (2015). Student perceptions of active learning. *College Student Journal*, 49(1), 121–133. Retrieved from https://www.ingentaconnect.com/content/prin/csj/2015/00000049/00000001/art00012

Lund, T., & Stains, M. (2015). The importance of context: an exploration of factors influencing the adoption of student-centered teaching among chemistry, biology, and physics faculty. *International Journal of STEM Education*, 2(1), 13. DOI:10.1186/s40594-015-0026-8

Lund, T., Pilarz, M., Velasco, J., Chakraverty, D., Rosploch, K., Undersander, M., & Stains, M. (2015). The best of both worlds: Building on the COPUS and RTOP observation protocols to easily and reliably measure various levels of reformed instructional practice. *CBE—Life Sciences Education*, 14(2), ar18, 1-12. DOI:10.1187/cbe.14-10-0168

Macpherson, R., & Stanovich, K. (2007). Cognitive ability, thinking dispositions, and instructional set as predictors of critical thinking. *Learning and Individual Differences,* 17, 115–127. DOI:10.1016/j.lindif.2007.05.003

Mahan, J. (1970). *Toward transdisciplinary inquiry in the humane sciences* (Doctoral dissertation, United States International University, San Diego, California). Retrieved from ProQuest Dissertations & Theses Global. (UMI 7020145).

Mahapoonyanont, N. (2010). Factors related to critical thinking abilities; a meta-analysis. *Procedia Social and Behavioral Sciences*, 9, 986-990. DOI:10.1016/j.sbspro.2010.12.272

Mahapoonyanont, N. (2012). The causal model of some factors affecting critical thinking abilities. *Procedia Social and Behavioral Sciences*, 46, 146-150. DOI:10.1016/j.sbspro.2012.05.084

Mahapoonyanont, N., Krahomwong, R., Kochakornjarupong, D., & Rachasong, W. (2010). Critical thinking abilities assessment tools: reliability generalization. *Procedia Social and Behavioral Sciences*, 2(2), 434-438. DOI:10.1016/j.sbspro.2010.03.038

Manalo, E., & Sheppard, C. (2016). How might language affect critical thinking performance? *Thinking Skills and Creativity,* 21, 41-49. DOI:10.1016/j.tsc.2016.05.005

Mangi, S., & Austen, M. (2008). Perceptions of stakeholders towards objectives and zoning of marine-protected areas in southern Europe. *Journal for Nature Conservation,* 16(4), 271-280. DOI:10.1016/j.jnc.2008.09.002

Marin, L., & Halpern, D. (2010). Pedagogy for developing critical thinking in adolescents: Explicit instruction produces greatest gains. *Thinking Skills and Creativity,* 6, 1-13. DOI:10.1016/j.tsc.2010.08.002

Marshall, J., & Alston, D. (2014). Effective, sustained inquiry-based instruction promotes higher science proficiency among all groups: A five-year analysis. *Journal of Science Teacher Education,* 25(7), 807-821. DOI:10.1007/s10972-014-9401-4

Marshall, J., Horton, R., & White, C. (2009). Equipping teachers. *The Science Teacher,* 76(4), 46. Retrieved from https://www.jstor.org/stable/24144731

Martineau, E., & Boisvert, L. (2011). Using Wikipedia to develop students' critical analysis skills in the undergraduate chemistry curriculum. *Journal of Chemical Education*, 88, 769-771. DOI:10.1021/ed100017k

Martinez, M. (2006). What is metacognition? *Phi Delta Kappan,* 87(9), 696-699. DOI:10.1177/003172170608700916

Mason, L., Ariasi, N., & Boldrin, A. (2011). Epistemic beliefs in action: Spontaneous reflections about knowledge and knowing during online information searching and their influence on learning. *Learning and Instruction*, 21, 137–151. DOI:10.1016/j.learninstruc.2010.01.001

Mathers, C., Finney, S., & Hathcoat, J. (2018). Student learning in higher education: A longitudinal analysis and faculty discussion. *Assessment & Evaluation in Higher Education,* 43(8), 1211-1227. DOI:10.1080/02602938.2018.1443202

Mayer, R., & Wittrock, M. (2006). Problem solving. In P. Alexander & P. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 287-304). Mahwah, New Jersey: Erlbaum.

Mayweg-Paus, E., Thiebach, M., & Jucks, R. (2016). Let me critically question this!–Insights from a training study on the role of questioning on argumentative discourse. *International Journal of Educational Research,* 79, 195-210. DOI:10.1016/j.ijer.2016.05.017

Mazer, J., Hunt, S., & Kuznekoff, J. (2007). Revising general education: Assessing a critical thinking instructional model in the basic communication course. *The Journal of General Education,* 56(3), 173-199. DOI:10.1353/jge.0.0000

McClellan, E. (2016). What a long, strange trip it's been: Three decades of outcomes assessment in higher education. *PS: Political Science & Politics*, 49(1).88-92. DOI:10.1017/S1049096515001298

McConnell, D., Chapman, L., Czajka, C., Jones, J., Ryker, K., & Wiggen, J. (2017). Instructional utility and learning efficacy of common active learning strategies. *Journal of Geoscience Education*, 65(4), 604–625. DOI:10.5408/17-249.1

McGinnis, D. (2016). Epistemological orientations and evidence evaluation in undergraduates. *Thinking Skills and Creativity,* 19, 279-289. DOI:10.1016/j.tsc.2016.01.002

McKendry, S. (2015). *Critical thinking skills for healthcare*. New York: Routledge.

McMichael, A. (2000). Transdisciplinarity in science. In M. Somerville, & D. Rapport (Eds.), *Transdisciplinarity: Recreating integrated knowledge*. Oxford: EOLSS Publishers Co.

McPeck, J. (1981). *Critical thinking and education*. Oxford: Martin Robertson.

McPeck, J. (1990). Critical thinking and subject specificity: A reply to Ennis. *Educational Researcher,* 19(4), 10-12. DOI:10.3102/0013189X019004010

Mercer-Mapstone, L., & Kuchel, L. (2017). Core skills for effective science communication: A teaching resource for undergraduate science education. *International Journal of Science Education, Part B*, 7(2), 181-201. DOI:10.1080/21548455.2015.1113573

Merrill, D. (2013). *First principles of instruction: Identifying and designing effective, efficient, and engaging instruction.* San Francisco: John Wiley & Sons, Inc.

Miele, D., & Wigfield, A. (2014). Quantitative and qualitative relations between motivation and critical-analytic thinking. *Educational Psychology Review,* 26, 519–541. DOI:10.1007/s10648-014-9282-2

Miller, D., & Czegan, D. (2016). Integrating the liberal arts and chemistry: A series of general chemistry assignments to develop science literacy. *Journal of Chemical Education*, 93(5), 864-869. DOI:10.1021/acs.jchemed.5b00942

Minner, D., Levy, A., & Century, J. (2009). Inquiry-based science instruction—what is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 47(4), 474-496. DOI:10.1002/tea.20347

Missett, T. (2012). *The development of critical and creative thinking skills for 21st century learning* (Doctoral dissertation, University of Virginia, Charlottesville, Virginia). Retrieved from ProQuest Dissertations & Theses Global. (UMI 3525032).

Mizell, S., & Brown, S. (2017). The current status of STEM education research 2013-2015. *Journal of STEM Education,* 17(4). Retrieved from https://www.learntechlib.org/p/174761/

Moffitt, T., Arseneault, L., Belsky, D., Dickson, N., Hancox, R., Harrington, H., Houts, R., Poulton, R., Roberts, B., Ross, S., Sears, M., Thomson, W., & Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences,* 108(7), 2693-2698.

Montuori, A. (2019). Creating social creativity: Integrative transdisciplinarity and the epistemology of complexity. In I. Lebuda & V. Glaveanu (Eds.), *The Palgrave handbook of social creativity* (pp. 407-430). Cham, Switzerland: Palgrave Macmillan.

Moon, J. (2004). *A handbook of reflective and experiential learning – theory and practice.* London: RoutledgeFalmer.

Moon, J. (2007). *Critical thinking: An exploration of theory and practice*. Abingdon, Oxon: Routledge.

Moore, T. (2013). Critical thinking: seven definitions in search of a concept. *Studies in Higher Education,* 38(4), 506-522. DOI:10.1080/03075079.2011.586995

Mosher, D. (2018). *The effect of mode of presentation, cognitive load, and individual differences on recall* (Doctoral dissertation, University of Reading, Berkshire, England). Retrieved from http://centaur.reading.ac.uk/84822/

Mouton, M., & Archer, E. (2018). Legitimation code theory to facilitate transition from high school to first-year biology. *Journal of Biological Education*, 53(1), 2-20. DOI:10.1080/00219 266.2017.1420681

Mueller, C., & Dweck, C. (2008). Praise for intelligence can undermine children's motivation and performance. *Journal of Personality and Social Psychology*, 75, 33-52. DOI:10.1037/0022-3514.75.1.33

Mulnix, J. (2012). Thinking critically about critical thinking. *Educational Philosophy and theory*, 44(5), 464-479. DOI:10.1111/j.1469-5812.2010.00673.x

Münnix, G. (2018). The method of thought experiments: A device to stimulate and develop critical thinking. *Proceedings of the XXIII World Congress of Philosophy*, 74, 43-50. DOI:10.5840/wcp232018741714

Murphy, S., MacDonald, A., Danaia, L., & Wang, C. (2019). An analysis of Australian STEM education strategies. *Policy Futures in Education,* 17(2), 122-139. DOI:10.1177/1478210318774190

Murray, N. (2001). *A history of apprenticeship in New Zealand* (Unpublished Master's Thesis). Lincoln University, Christchurch, New Zealand. Retrieved from http://hdl.voced.edu.au/10707/253539

Nagi, K. (2018, July). New social media and impact of fake news on society. *ICSSM Proceedings*, Chaing Mai, Thailand, 77-96. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3258350

National Academy of Science (2009). *Strengthening high school chemistry education through teacher outreach program: A workshop summary to the Chemical Sciences Roundtable*. Washington, DC: The National Academies Press. DOI:10.17226/12533

National Defence Education Act of 1958. (1958). Public Law 85-864, 85th Congress. 70 STAT 188. (enacted).

National Institute of Education (1984). *Involvement in learning: Realizing the potential of American higher education.* Washington, DC: National Institute of Education, U.S. Department of Education.

National Research Council (2003). *BIO 2010: Transforming undergraduate education for future research biologists*, Washington, DC: National Academies Press.

Nevid, J., Ambrose, M., & Pyun, Y. (2017). Effects of higher and lower level writing-to-learn assignments on higher and lower level examination questions. *Teaching of Psychology*, 44(4), 324–329. DOI:10.1177/0098628317727645

New Jersey Minority Educational Development (2019). *World top 20 project*. Accessed 30/03/2019 at https://worldtop20.org/education-data-base

Nicholas, M. (2011). *Faculty conceptualizations and approaches to assessing critical thinking in the humanities and natural Sciences—A grounded theory study* (Doctoral dissertation, University of Cincinnati, Cincinnati, Ohio). Retrieved from ProQuest Dissertations & Theses Global. (UMI 3491385).

Nicholas, M., & Labig, C. (2013). Faculty approaches to assessing critical thinking in the humanities and the natural and social sciences: Implications for general education. *The Journal of General Education*, 62(4), 297-319. DOI:10.1353/jge.2013.002

Nicholas, M., & Raider-Roth, M. (2016). A hopeful pedagogy to critical thinking. *International Journal for the Scholarship of Teaching and Learning*, 10(2), Article 3, 1-10. DOI:10.20429/ijsotl.2016.100203

Niu, L., Behar-Horenstein, L., & Garvan, C. (2013). Do instructional interventions influence college students critical thinking skills? a meta-analysis. *Educational Research Review*, 9, 114-128. DOI:10.1016/j.edurev.2012.12.002

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies create the dynamics of innovation.* Oxford, England: Oxford University Press.

Norris, S. (1985). Synthesis of research on critical thinking. *Educational Leadership*, 42, 40-45.

Norris, S. (1989). Can we test validly for critical thinking? *Educational Researcher*, 18(9), 21-26. DOI:10.3102/0013189X018009021

Norris, S. (1992). Introduction: The generalizability question. In S. Norris (Ed.), *The generalizability of critical thinking: Multiple perspectives on an educational ideal* (pp. 1–8). New York: Teachers College Press.

Nosich, G. (1996). The need for comprehensiveness in critical thinking instruction. *Inquiry: Critical Thinking Across the Disciplines*, 16(2), 50-66. DOI:10.5840/inquiryctnews199616231

Noubani, A. (2018). *Improving faculty professional development for community college science educators* (Doctoral dissertation, University of Delaware, Newark, Delaware). Retrieved from ProQuest Dissertations & Theses Global. (UMI 10978073).

Novak, J. (1990). Concept mapping: A useful tool for science education. *Journal of Research in Science Teaching,* 27, 937–949. DOI:10.1002/tea.3660271003

Nusche, D. (2008). Assessment of learning outcomes in higher education: A comparative review of selected practices. *OECD Education Working Papers*, 15. Paris: Organisation for Economic Co-operation and Development Publishing. DOI:10.1787/244257272573

O'Hare, L., & McGuinness, C. (2015). The validity of critical thinking tests for predicting degree performance: A longitudinal study. *International Journal of Educational Research,* 72, 162-172. DOI:10.1016/j.ijer.2015.06.004

Oliver, B. (2011). *Assuring graduate outcomes good practice report.* Sydney, New South Wales: Australian Learning and Teaching Council.

Oliver, B., & Jorre de St Jorre, T. (2018). Graduate attributes for 2020 and beyond: recommendations for Australian higher education providers. *Higher Education Research & Development,* 37(4), 821-836. DOI:10.1080/07294360.2018.1446415

Oliver-Hoyo, M. (2003). Designing a written assignment to promote the use of critical thinking skills in an introductory chemistry course. *Journal of Chemical Education*, 80(8), 899. DOI:10.1021/ed080p899

Oliver-Hoyo, M., Allen, D., & Anderson, M. (2004). Inquiry-guided instruction: practical issues of implementation. *Journal of College Science Teaching,* 33(6), 20-24. Retrieved from https://www.jstor.org/stable/26491298

Oliveri, M., Lawless, R., & Young, J. (2015) *A validity framework for the use and development of exported assessments.* Princeton, New Jersey: Educational Testing Service. Retrieved from https://www.ets.org/s/about/pdf/exported_assessments.pdf

Oral, O., Tanriöver, Ö., & Soubra, M. (2017). Modeling and predicting scientific thinking skills of university students using a data mining tool. *International Journal of Informatics Technologies*, 10(1), 89-98. DOI:10.17671/btd.25506

Organisation for Economic Co-operation and Development (2013a). *Assessment of Higher Education Learning Outcomes feasibility study report*. *Volume 2 - data analysis and* *national experiences.* Paris: Organisation for Economic Co-operation and Development Publishing. Retrieved from http://www.oecd.org/education/skills-beyond-school/AHELOFSReportVolume2.pdf

Organisation for Economic Co-operation and Development (2013b). *Assessment of Higher Education Learning Outcomes feasibility study report. Volume 3 - further insights*. Paris: Organisation for Economic Co-operation and Development Publishing. Retrieved from http://www.oecd.org/education/skills-beyond-school/AHELOFSReportVolume3.pdf

Organisation for Economic Co-operation and Development (2014). *PISA 2012 Results: Creative problem solving. Volume V - Students' skills in tackling real-life problems*. Paris: Organisation for Economic Co-operation and Development Publishing. Retrieved from https://www.oecd.org/education/pisa-2012-results-volume-v.htm

Organisation for Economic Co-operation and Development (2018). PISA 2015 results in focus. *PISA in Focus,* 67, Paris: Organisation for Economic Co-operation and Development Publishing. DOI:10.1787/22260919

Ornstein, A., & Hunkins, F. (2004). *Curriculum: Foundations, principles and issues* (3rd ed). Boston: Allyn and Bacon.

Ortiz, C. (2007). *Does philosophy improve critical thinking skills?* (Master's thesis, University of Melbourne, Victoria, Australia). Retrieved from https://www.reasoninglab.com/wp-content/uploads/2015/10/Alvarez-Final_Version.pdf

Osborne, R., Kriese, P., Tobey, H., & Johnson, E. (2009). Putting it all together: Incorporating "SoTL practices" for teaching interpersonal and critical thinking skills in an online course. *InSight: A Journal of Scholarly Teaching*, 4, 45-55. Retrieved from https://files.eric.ed.gov/fulltext/EJ864285.pdf

Palmer, S., Campbell, E. Johnson, & West, J. (2017). Occupational outcomes for bachelor of science graduates in Australia and implications for undergraduate science curricula. *Research in Science Education,* 48(5), 989–1006. DOI:10.1007/s11165-016-9595-x

Paris, S., & Winograd, P. (1990). How metacognition can promote academic Learning and Instruction, In B. Jones & L. Idol (Eds.), *Dimensions of thinking and cognitive instruction* (pp. 15-51). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Partin, M., Underwood, E., & Worch, E. (2013). Factors related to college students' understanding of the nature of science: comparison of science majors and nonscience majors. *Journal of*

College Science Teaching, *42*(6), 89-99. Retrieved from https://www.jstor.org/stable/43632161

Pascarella, E. (2006). How college affects students: Ten directions for future research. *Journal of College Student Development,* 47(5), 508-520. DOI:10.1353/csd.2006.0060

Pascarella, E., & Terenzini, P. (1991). *How college affects students: findings and insights from twenty years of research.* San Francisco, California: Jossey-Bass Publishers.

Pascarella, E., & Terenzini, P. (2005). *How college affects students: a third decade of research.* San Francisco, California: Jossey-Bass Publishers.

Paul, R. (1985). The critical thinking movement: A historical perspective. *National Forum: Phi Kappa Phi Journal,* 65(1), 2-3. Retrieved from https://search-proquest-com.ezproxy.flinders.edu.au/docview/63385466?accountid=10910

Paul, R. (1992). Critical thinking: What, why, and how? *New Directions for Community Colleges*, 77, 3–24. DOI:10.1002/cc.36819927703

Paul, R. (1995). *Socratic questioning and role-playing*. Santa Rosa, California: Foundation for Critical Thinking.

Paul, R. (1996). Critical thinking and the state of education today. *Inquiry: Critical Thinking Across the Disciplines*, 16(2), 12-34. DOI:10.5840/inquiryctnews199616214

Paul, R. (2005). The state of critical thinking today. *New Directions for Community Colleges*, 130, 27-38. DOI:10.1002/cc.193

Paul, R. (2011). Reflections on the nature of critical thinking, its history, politics, and barriers and on its status across the college/university curriculum: Part I. *Inquiry: Critical Thinking Across the Disciplines*, 26(3), 5-24. DOI:10.5840/inquiryct201126324

Paul, R. (2012). Reflections on the nature of critical thinking, its history, politics, and barriers and on its status across the college/university curriculum: Part 2. *Inquiry: Critical Thinking Across the Disciplines*, 27(1), 5-30. DOI:10.5840/inquiryct20122712

Paul, R., & Elder, L. (2002). *Critical thinking: Tools for taking charge of your professional and personal life*. Upper Saddle River, New Jersey: Pearson Education/FT Press.

Paul, R., & Elder, L. (2006a). *The miniature guide to critical thinking: Concepts and tools.* Foundation for Critical Thinking. Last viewed 21 March 2018, https://www.criticalthinking.org/files/Concepts_Tools.pdf

Paul, R., & Elder, L. (2006b). Paul-Elder framework [diagram]. *The Miniature Guide to Critical Thinking: Concepts and Tools.* Foundation for Critical Thinking. page 25.

Paul, R. & Elder, L. (2008). Critical thinking: The art of Socratic questioning. Journal of Developmental Education, 31(2), 32-33. Retrieved from https://files.eric.ed.gov/fulltext/EJ832681.pdf

Paul, R., Binker, A., Martin, D., & Adamson, K. (1989). *Critical thinking handbook: High school. A guide for redesigning instruction.* Rohnert Park, California: Center for Critical Thinking and Moral Critique, Sonoma State University.

Paul, R., Elder, L., & Bartell, T. (1997a). *California teacher preparation for the instruction of critical thinking: Research findings and recommendations*. Sacramento: California Commission on Teaching Credentialing.

Paul, R., Elder, L., & Bartell, T. (1997b). *Study of 38 public universities and 28 private universities to determine faculty emphasis on critical thinking in instruction*. Tomales, California: Foundation for Critical Thinking. Retrieved from http://www.criticalthinking.org/pages/study-of-38-public-universities-and-28-private-universities-to-determine-faculty-emphasis-on-critical-thinking-in-instruction/598

Paul, R., Fisher, A., & Nosich, G. (1993). *Workshop on critical thinking strategies*. Sonoma State University, California: Foundation for Critical Thinking.

Pearl, A., Rayner, G., Larson, I., & Orlando, L. (2019). Thinking about critical thinking: An industry perspective. *Industry and Higher Education*, 33(2), 116-126. DOI:10.1177/0950422218796099

Peeters, M. (2016). Practical significance: moving beyond statistical significance. *Currents in Pharmacy Teaching and Learning,* 8(1), 83-89. DOI:10.1016/j.cptl.2015.09.001

Peeters, M., & Boddu, S. (2016). Assessing development in critical thinking: One institution’s experience. *Currents in Pharmacy Teaching and Learning*, 8(3), 271-278. DOI:10.1016/j.cptl.2016.02.007

Peirce, C. (1877). The fixation of belief. *The Popular Science Monthly*, 12, 1-15. Retrieved from https://www.biodiversitylibrary.org/item/18054#page/11/mode/1up

Pellegrino, J., & Hilton, M. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: National Academy of Sciences

Perkins, D. (1995). *Outsmarting IQ*: *The emerging science of learnable intelligence*. New York: Free Press.

Perkins, D., Jay, E., & Tishman, S. (1993). Beyond abilities: A dispositional theory of thinking. *Merrill-Palmer Quarterly*, 39, 1-21. Retrieved from http://www.jstor.org/stable/23087298

Petek, E., & Bedir, H. (2018). An adaptable teacher education framework for critical thinking in language teaching. *Thinking Skills and Creativity,* 28, 56-72. DOI:10.1016/j.tsc.2018.02.008

Phelan, J. (2012). *A teacher action research study: enhancing student critical thinking knowledge, skills, dispositions, application and transfer in higher education technology course* (Doctoral dissertation, Capella University, Minneapolis, Minnesota). Retrieved from ProQuest Dissertations & Theses Global. (UMI 3549444).

Phillips, F., & Mackintosh, B. (2011). Wiki art gallery Inc.: A case for critical thinking. *Issues in Accounting Education*, 26(3), 593–608. DOI:10.2308/iace-50038

Pietrzak, B., Ward, A., Cheung, M., Schwendimann, B., Mollaoglu, G., Duong, M., Ulltveit-Moe, N. Allareddy, V., Dutton-Regester, K., Zhang, J., Scult, M., Naz, S., Singh, P., Yan, H., Isaacson, K., Dennis, A., Al-Humaidan, E., Beardsley, F., Lo, C., Sood, P., Jones, T., Nieuwenhuis, R., Ali, B., Yu, K., Arthur, P., Kumar, B., Chen, A., Buschke, F., Cingl, L., Zaidi, S., O'mullane, A., Coetzee, V., & Konstantinides, N. (2018). Education for the future. *Science*, 360(6396), 1409-1412. DOI:10.1126/science.aau3877

Pintrich, P., Marx, R., & Boyle, R. (1993). Beyond cold conceptual change: the role of motivational beliefs and classroom contextual factors in the process of conceptual change. *Review of Educational Research*, 63, 167–199. DOI:10.3102/00346543063002167

Pithers, R., & Soden, R. (2000). Critical thinking in education: A review. *Educational Research*, 42(3), 237–249. DOI:10.1080/001318800440579

Pitman, T., & Broomhall, S. (2009). Australian universities, generic skills and lifelong learning. *International Journal of Lifelong Education*, 28(4), 439-458 DOI:10.1080/02601370903031280

Plotnikova, N., & Kondrateva, I. (2018). Development of integrative skills in higher education students. *Helix*, 8(1). 2877-2883. DOI:10.29042/2018-2877-2883

Pnevmatikos, D., Christodoulou, P., & Georgiadou, T. (2019). Promoting critical thinking in higher education through the values and knowledge education (VaKE) method. *Studies in Higher Education,* 44(5), 892-901. DOI:10.1080/03075079.2019.1586340

Poirier, G., & Hocker, S. (1993). Teaching critical thinking in a library credit course. *Research Strategies,* 11(4), 233-41. Retrieved from ERIC (EJ479762).

Polomba, C., & Banta, T. (1999). *Assessment essentials.* San Francisco, California: Jossey-Bass Publishers.

Porter, S. (2013). Self-reported learning gains: A theory and test of college student survey response. *Research in Higher Education,* 54(2), 201-226. DOI:10.1007/s11162-012-9277-0

Possin, K. (2002). *Critical thinking.* Winona, Minnesota: The Critical Thinking Lab.

Possin, K. (2008). A field guide to critical-thinking assessment. *Teaching Philosophy,* 21(3), 201-228. DOI:10.5840/teachphil200831324

Postareff, L., Mattsson, M., & Parpala, A. (2018). The effect of perceptions of the teaching-learning environment on the variation in approaches to learning–Between-student differences and within-student variation. *Learning and Individual Differences,* 68, 96-107. DOI:10.1016/j.lindif.2018.10.006

Power, E., & Handley, J. (2019). A best-practice model for integrating interdisciplinarity into the higher education student experience. *Studies in Higher Education,* 44(3), 554-570. DOI:10.1080/03075079.2017.1389876

Preacher, K., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods,* 16(2), 93–115. DOI:10.1037/a0022658

Preiss, D., Castillo, J., Flotts, P., & San Martín, E. (2013). Assessment of argumentative writing and critical thinking in higher education: Educational correlates and gender differences. *Learning and Individual Differences*, 28, 193–203. DOI:10.1016/j.lindif.2013.06.004

Pressley, M., Wharton-McDonald, R., Mistretta-Hampton, J., & Echevarria, M. (1998). Literacy instruction in 10 fourth-and fifth-grade classrooms in upstate New York. *Scientific Studies of Reading,* 2(2), 159–194. DOI:10.1207/s1532799xssr0202_4

Prince, M., & Felder, R. (2007). The many faces of inductive teaching and learning. *Journal of College Science Teaching,* 36(5), 14-20. Retrieved from https://www.jstor.org/stable/42992681

Prince, M., Felder, R., & Brent, R. (2007). Does faculty research improve undergraduate teaching? an analysis of existing and potential synergies. *Journal of Engineering Education,* 96(4), 283-294. DOI:10.1002/j.2168-9830.2007.tb00939.x

Prinsley, R., & Baranyai, K. (2015). STEM skills in the workforce: What do employers want? *Occasional Paper Series*, Issue 9, 1-4. Canberra, Australia: Office of the Chief Scientist.

Puig, B., Blanco-Anaya, P., Bargiel, I., & Crujeiras-Pérez, B. (2019). A systematic review on critical thinking intervention studies in higher education across professional fields. *Studies in Higher Education,* 1-10. DOI:10.1080/03075079.2019.1586333

Puolimatka, T. (2003). Constructivism and critical thinking. *Inquiry: Critical Thinking Across the Disciplines*, 22(4), 5-12. DOI:10.5840/inquiryctnews20032244

Qing, Z., Ni, S., & Hong, T. (2010). Developing critical thinking disposition by task-based learning in chemistry experiment teaching. *Procedia Social and Behavioral Sciences*, 2, 4561-4570. DOI:10.1016/j.sbspro.2010.03.731

Quitadamo, I., & Kurtz, M. (2007). Learning to improve: Using writing to increase critical thinking performance in general education biology. *CBE-Life Sciences Education,* 6, 140-154. DOI:10.1187/cbe.06–11–0203

Rabu, S., Aris, B., & Tasir, Z. (2013). Teaching critical thinking through online instructor scaffolding: A conceptual framework. *Procedia Social and Behavioral Sciences*, 97, 314-319. DOI:10.1016/j.sbspro.2013.10.239

Ramsden, J. (1997). How does a context-based approach influence understanding of key chemical ideas at 16+? *International Journal of Science Education*, 19(6), 697–710. DOI:10.1080/0950069970190606

Rapanta, C., & Macagno, F. (2016). Argumentation methods in educational contexts: Introduction to the special issue. *International Journal of Educational Research,* 79, 142-149. DOI:10.1016/j.ijer.2016.07.008

Rayner, G., & Papakonstantinou, T. (2018). Foundation biology students' critical thinking ability: self-efficacy versus actuality. *Journal of University Teaching and Learning Practice*, 15(5), 4. Retrieved from https://ro.uow.edu.au/jutlp/vol15/iss5/4

Rear, D. (2019). One size fits all? The limitations of standardised assessment in critical thinking, *Assessment & Evaluation in Higher Education*, 44(5), 664-675. DOI:10.1080/02602938.2018.1526255

Reece, G. (2002). *Critical thinking and transferability: A review of the literature*. Retrieved from http://people.umass.edu/curtis/academics/researchtoolbox/pdfs/Reece_LitReview_2010.pdf

Reece, G. (2005). Critical thinking and cognitive transfer: Implications for the development of online information literacy tutorials. *Research Strategies*, 20(4), 482-493. DOI:10.1016/j.resstr.2006.12.018

Renaud, R., & Murray, H. (2007). The validity of higher-order questions as a process indicator of educational quality. *Research in Higher Education,* 48(3), 319-351. DOI:10.1007/s11162-006-9028-1

Renaud, R., & Murray, H. (2008). A comparison of a subject-specific and a general measure of critical thinking. *Thinking Skills and Creativity,* 3, 85-93. DOI:10.1016/j.tsc.2008.03.005

Repo, S., Lehtinen, T., Rusanen, E., & Hyytinen, H. (2017). Prior education of Open University students contributes to their capability in critical thinking. *Journal of Adult and Continuing Education*, 23(1), 61-77. DOI:10.1177/1477971417693416

Reynolds, S. (2016). *Determining and exploring teachers' perceptions on the barriers to teaching critical thinking in the classroom: A survey study* (Doctoral dissertation, Texas Tech University, Lubbock, Texas). Retreived from https://ttu-ir.tdl.org/handle/2346/68118

Rhoades, E., Ricketts, J., & Friedel, C. (2008). Major comparison of cognitive potential: Are agriculture students different? *Proceedings of the 2008 American Association of Agricultural Education National Research Conference*, 35, 165-179. Retrieved from http://aaaeonline.org/Resources/Documents/Research%20Conference%20Proceedings,%20North%20Central%202008.pdf

Rickert, R. (1967). Developing critical thinking. *Science Education*, 51(1), 24-27. DOI:10.1002/sce.3730510106

Rieh, S., & Hilligoss, B. (2008). College students' credibility judgments in the information-seeking process. Digital media, youth, and credibility. In M. Metzger & A. Flanagin (Eds.), *The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning* (pp. 49–72). Cambridge, Massachusetts: The MIT Press.

Rios, J., Guo, H., Mao, L., & Liu, O. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing*, 17(1), 74-104. DOI:10.1080/15305058.2016.1231193

Roberts, D. (2007). Scientific literacy. In S. Abell & N. Lederman (Eds.), *Handbook of research on science education* (pp. 125–171). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Roberts, D. (2019). Higher education lectures: From passive to active learning via imagery? *Active Learning in Higher Education,* 20(1), 63-77. DOI:10.1177/1469787417731198

Robinson, S. (2011). Teaching logic and teaching critical thinking: revisiting McPeck. *Higher Education Research and Development,* 30(3), 275-287. DOI:10.1080/07294360.2010.500656

Roehrig, G., Michlin, M., Schmitt, L., MacNabb, C., & Dubinsky, J. (2012). Teaching neuroscience to science teachers: Facilitating the translation of inquiry-based teaching instruction to the classroom. *CBE-Life Sciences Education,* 11(4), 413-424. DOI:10.1187/cbe.12-04-0045

Roohr, K., Liu, H., & Liu, O. (2017). Investigating student learning gains in college: a longitudinal study. *Studies in Higher Education,* 42(12), 2284-2300. DOI:10.1080/03075079.2016.1143925

Rose, L., Gillespie, M., Rowe, M., Primm, P., & Shannon, L. (2012). Extraordinary claims: An innovative approach to engage student interest and enhance critical thinking skills in general education science courses. In K. McMahon (Ed.), *Tested studies for laboratory teaching. Proceedings of the 33rd Conference of the Association for Biology Laboratory Education* (Vol. 33, pp. 311-313). Retrieved from http://www.ableweb.org/biologylabs/wp-content/uploads/volumes/vol-33/34_Rose.pdf

Roksa, J., & Arum, R. (2011). The state of undergraduate learning. *Change: The Magazine of Higher Learning,* 43(2), 35–38. DOI:10.1080/00091383.2011.556992

Roksa, J., Trolian, T., Blaich, C., & Wise, K. (2017). Facilitating academic performance in college: understanding the role of clear and organized instruction. *Higher Education*, 74(2), 283-300. DOI:10.1007/s10734-016-0048-2

Rutledge, M. (2005). Making the nature of science relevant: Effectiveness of an activity that stresses critical thinking. *The American Biology Teacher,* 67(6), 329-333. DOI:10.2307/4451857

Rutledge, M., Bonner, J., & Lampley, S. (2015). The impact of active-learning exercises on the grade distribution in a large-lecture, general education biology course. *Journal of College Science Teaching*, 44(5), 16-23. Retrieved from https://www.jstor.org/stable/43631843

Saadati, F., Tarmizi, R., & Bayat, S. (2010). Assessing critical thinking of postgraduate students. *Procedia Social and Behavioral Sciences*, 8, 543-548. DOI:10.1016/j.sbspro.2010.12.075

Saefi, M., Suwono, H., & Susilo, H. (2016, November). *Biology student teacher's critical thinking: an exploration study.* Paper presented at International Conference on Education: Education in the 21th Century: Responding to Current Issues**,** Malang, East Java, Indonesia (pp. 605-612). Retrieved from http://pasca.um.ac.id/conferences/index.php/ice/article/view/82

Saeverot, H., & Kvam, V. (2019). An alternative model of researching educational practice: A pedagogic–stereoscopic point of view. *British Educational Research Journal*, 45(1), 201-218. DOI:10.1002/berj.3493

Saido, G., Siraj, S., Nordin, A., & Al Almedy, O. (2015). Higher order thinking skills among secondary school students in science learning. *MOJES: Malaysian Online Journal of Educational Sciences,* 3(3), 13-20. Retrieved from: http://mojes.um.edu. my/index.php/MOJES/article/view/12778

Sanders, M., & Moulenbelt, J. (2011). Defining critical thinking. How far have we come? *Inquiry: Critical Thinking Across the Disciplines*, 26(1), 38-46. DOI:10.5840/inquiryctnews20112616

Sandmire, D., & Boyce, P. (2004). Pairing of opposite learning styles among allied health students: Effects on collaborative performance. *Journal of Allied Health*, 33(2), 156-163. Retrieved from https://search.proquest.com/docview/210979341?accountid=10910

Sarican, G., & Akgunduz, D. (2018). The impact of integrated STEM Education on academic achievement, reflective thinking skills towards problem solving and permanence in learning in science education. *Cypriot Journal of Educational Sciences*, 13(1), 94-107. DOI:10.18844/cjes.v13i1.3372

Sarigoz, O. (2012). Assessment of the high school students' critical thinking skills. *Procedia Social and Behavioral Sciences*, 46, 5315-5319. DOI:10.1016/j.sbspro.2012.06.430

Sasson, I., Yehuda, I., & Malkinson, N. (2018). Fostering the skills of critical thinking and question-posing in a project-based learning environment. *Thinking Skills and Creativity,* 29, 203-212. DOI:10.1016/j.tsc.2018.08.001

Schmaltz, R., Jansen, E., & Wenckowski, N. (2017). Redefining critical thinking: Teaching students to think like scientists. *Frontiers in Psychology,* 8, Article 459, 1-4. DOI:10.3389/fpsyg.2017.00459

Schneider, C., & Miller, R. (2005). *Liberal education outcomes: A preliminary report on student achievement in college.* Washington DC: Association of American Colleges and Universities. Retrieved from http://www.aacu.org/sites/default/files/files/LEAP/LEAP_Report_2005.pdf

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science,* 26, 113–125. DOI:10.1023/A:1003044231033

Schraw, G., & Dennison, R. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology*, 19, 460-475. DOI:10.1006/ceps.1994.1033

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review*, 7(4), 351-371. DOI:10.1007/BF02212307

Schraw, G., Crippen, K., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education*, 36, 111-139. DOI:10.1007/s11165-005-3917-8

Sellars, M., Fakirmohammad, R., Bui, L., Fishetti, J., Niyozov, S., Reynolds, R. Thapliyal, N., Liu-Smith, Y., & Ali, N. (2018). Conversations on critical thinking: Can critical thinking find its way forward as the skill set and mindset of the century? *Education Sciences*, 8(4), 205. DOI:10.3390/educsci8040205

Sessoms, J., & Finney, S. (2015). Measuring and modeling change in examinee effort on low-stakes tests across testing occasions. *International Journal of Testing*, 15(4), 356-388. DOI:10.1080/15305058.2015.1034866

Seymour, E., Wiese, D., Hunter, A., & Daffinrud, S. (2000, March). *Creating a better mousetrap: On-line student assessment of their learning gains.* Paper presented at National Meetings of the American Chemical Society Symposium, San Francisco, California. Retrieved from https://salgsite.net/docs/SALGPaperPresentationAtACS.pdf

Sgambi, L., Kubiak, L., Basso, N., & Garavaglia, E. (2019). Active learning for the promotion of students' creativity and critical thinking. *Archnet-IJAR*, 13(2), 386-407. DOI:10.1108/ARCH-11-2018-0018

Shadle, S., Marker, A., & Earl, B. (2017). Faculty drivers and barriers: laying the groundwork for undergraduate STEM education reform in academic departments. *International Journal of STEM Education*, 4(1). DOI:10.1186/s40594-017-0062-7

Shavelson, R., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., & Marino, J. (2019). Assessment of university students' critical thinking: Next generation performance assessment. *International Journal of Testing*, 19(4) 337-362. DOI:10.1080/15305058.2018.1543309

Shehab, H., & Nussbaum E. (2015). Cognitive load of thinking strategies. *Learning and Instruction,* 35, 51-61. DOI:10.1016/j.learninstruc.2014.09.004

Sheldrake, R., Mujtaba, T., & Reiss, M. (2017). Science teaching and students' attitudes and aspirations: The importance of conveying the applications and relevance of science. *International Journal of Educational Research,* 85, 167-183. DOI:10.1016/j.ijer.2017.08.002

Shell, R. (2001). Perceived barriers to teaching for critical thinking by BSN nursing faculty. *Nursing and Health Care Perspectives*, 22(6), 286-291. Retrieved from https://search.proquest.com/docview/236596828?accountid=10910

Shillady, A. (2011). Fostering critical thinking and problem-solving skills in young children. *Young Children*, 66(5), 12-13. Retrieved from https://search.proquest.com/docview/896478630?accountid=10910

Shin, H., Park, C., & Kim, H. (2015). Validation of Yoon's critical thinking disposition instrument. *Asian nursing research*, 9(4), 342-348. DOI:10.1016/j.anr.2015.10.004

Shipley, T., McConnell, D., McNeal, K., Petcovic, H., & John, K. (2017). Transdisciplinary science education research and practice: Opportunities for GER in a developing STEM discipline-based education research alliance (DBER-A). *Journal of Geoscience Education*, 65(4), 354-362.

Shively, K., Stith, K., & Rubenstein, L. (2018). Measuring what matters: Assessing creativity, critical thinking, and the design process. *Gifted Child Today*, 41(3), 149-158. DOI:10.1177/1076217518768361

Shulman, L. (2000). From Minsk to Pinsk: Why a scholarship of teaching and learning? *Journal of Scholarship of Teaching and Learning*, 1, 48–53. Retrieved from https://scholarworks.iu.edu/journals/index.php/josotl/article/view/1582

Shweder, R. (2003). *Why do men barbecue?: Recipes for cultural psychology*. Cambridge, Massachusetts: Harvard University Press.

Siegel, H. (1988). *Educating reason*. New York: Routledge.

Siegel, M., Roberts, T., Freyermuth, S., Witzig, S., & Izci, K. (2015). Aligning assessment to instruction: collaborative group testing in large-enrolment science classes. *Journal of College Science Teaching,* 44(6), 74-82. Retrieved from https://www.jstor.org/stable/43631999

Silvia, P., & Sanders, C. (2010). Why are smart people curious? Fluid Intelligence, openness to experience, and interest. *Learning and Individual Differences,* 20, 242-245. DOI:10.1016/j.lindif.2010.01.006

Simper, N., Frank, B., Kaupp, J., Mulligan, N., & Scott, J. (2019). Comparison of standardized assessment methods: logistics, costs, incentives and use of data. *Assessment & Evaluation in Higher Education*, 44(6), 821-834. DOI:10.1080/02602938.2018.1533519

Singer, S., Nielsen, N., & Schweingruber, H. (2012). *Discipline-based education research: Understanding and improving learning in undergraduate science and engineering*. Washington, DC: National Academies Press.

Smith, M., Jones, F., Gilbert, S., & Wieman, C. (2013). The Classroom Observation Protocol for Undergraduate STEM (COPUS): a new instrument to characterize university STEM classroom practices. *CBE-Life Sciences Education*, 12(4), 618-627. DOI:10.1187/cbe.13-08-0154

Smith, M., Vinson, E., Smith, J., Lewin, J., & Stetzer, M. (2014). A campus-wide study of STEM courses: new perspectives on teaching practices and perceptions. *CBE-Life Sciences Education*, 13(4), 624-635. DOI:10.1187/cbe.14-06-0108

Smith, R. (2002). *Challenging your preconceptions: Thinking critically about psychology.* Belmont, California: Wadsworth/Thomson Learning.

Smith, T., Rama, P., & Helms, J. (2018). Teaching critical thinking in a ge class: A flipped model. *Thinking Skills and Creativity,* 28, 73-83. DOI:10.1016/j.tsc.2018.02.010

Smith, W. (2008). *Does gender influence online survey participation?: A record-linkage analysis of university faculty online survey response behavior.* Research Report. San Jose, California: San Jose State University. https://files.eric.ed.gov/fulltext/ED501717.pdf

Smyth, T. (2004). *The principles of writing in psychology*. London: Red Globe Press.

Solanki, V. (2018). *An evaluation of high stakes standardized tests and teacher instruction on high school student 21st century skills* (Doctoral dissertation, Northcentral University, San Diego, California). Retrieved from ProQuest Dissertations & Theses Global. (UMI 10785375).

Somerville, M., & Rapport, D. (Eds.). (2002). *Transdisciplinarity: Recreating integrated knowledge*. Oxford: EOLSS Publishers Co.

Sosu, E. (2012). The development and psychometric validation of a Critical Thinking Disposition Scale. *Thinking Skills and Creativity,* 9, 107-119. DOI:10.1016/j.tsc.2012.09.002

South Australian Department of Education and Children's Services (2010). *South Australian teaching for effective learning framework guide: a resource for developing quality teaching and learning in South Australia*. Dept. of Education and Children's Services, South Australia: Lane Print and Post.

Springer, C., & Borthick, A. (2004). Business Simulation to stage critical thinking in introductory accounting: Rationale, design, and implementation. *Issues in Accounting Education*, 19(3), 277-303. DOI:10.2308/iace.2004.19.3.277

Staib, S. (2003). Teaching and measuring critical thinking. *Journal of Nursing Education*, 42(11), 498–508. DOI:10.3928/0148-4834-20031101-08

Stains, M., Harshman, J. Barker, M., Chasteen, S., Cole, R., DeChenne-Peters, S., Eagan Jr., M., Esson, J., Knight, J., Laski, F., Levis-Fitzgerald, M., Lee, C., Lo, S., McDonnell, L., McKay, T., Michelotti, N., Musgrove, A., Palmer, M., Plank, K., Rodela, T., Sanders, E., Schimpf, N., Schulte, P., Smith, M., Stetzer, M., Van Valkenburgh, B., Vinson, E., Weir, L., Wendel, P., Wheeler, L., & Young, A. (2018). *Anatomy of STEM teaching in North American universities. Science,* 359(6383), 1468–1470. DOI:10.1126/science.aap889

Stanton, J., Neider, X., Gallegos, I., & Clark, N. (2015). Differences in metacognitive regulation in introductory biology students: when prompts are not enough. *CBE-Life Sciences Education,* 14(2), ar15. DOI:10.1187/cbe.14-08-0135

Stedman, N., & Adams, B. (2012). Identifying faculty's knowledge of critical thinking concepts and perceptions of critical thinking instruction in higher education. *NACTA Journal*, 56(2), 9-14. Retrieved from http://www.jstor.org/stable/nactajournal.56.2.9

Steger, M. (2019). Committing to cultures of creativity: the significance of transdisciplinarity. *Globalizations*, 16(5), 763-769. DOI:10.1080/14747731.2019.1578485

Stein, B., & Haynes, A. (2011) Engaging faculty in the assessment and improvement of students' critical thinking using Critical thinking Assessment Test. *Change: The Magazine of Higher Learning*, 43(2), 44-49. DOI:10.1080/00091383.2011.550254

Stein, B., Haynes, A., & Redding, M. (2006). *Project CAT: Assessing critical thinking skills year two report*. Tennessee Technological University, Cookeville, Tennessee. Retrieved from http://www.tntech.edu/cat

Stein, B., Haynes, A., & Unterstein, J. (2003, December). *Assessing critical thinking skills*. Paper presented at Southern Association of Colleges and Schools Commission on Colleges Annual Meeting, Nashville, Tennessee.

Stein, B., Haynes, A., Redding, M., Ennis, T., & Cecil, M. (2007). Assessing critical thinking in STEM and beyond. In M. Iskander (Ed.), *Innovations in e-learning, instruction technology, assessment, and engineering education* (pp. 79-82). Dordrecht, Netherlands: Springer.

Stephenson, N., & Sadler-Mcknight, N. (2016). Developing critical thinking skills using the science writing heuristic in the chemistry laboratory. *Chemistry Education Research and Practice*, 17(1), 72-79. DOI:10.1039/C5RP00102A

Sternberg, R. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.

Sternberg, R. (1986). *Critical thinking: Its nature, measurement, and improvement*. Washington, DC: National Institute of Education.

Sternberg, R. (1987). Questions and answers about the nature and teaching of thinking skills. In J. Baron and R. Sternberg (Eds.), *Teaching thinking skills: Theory and practice* (pp. 251-259). New York: Freeman.

Sternberg, R. (1997). The concept of intelligence and its role in lifelong learning and success. *American Psychologist,* 52(10), 1030. DOI:10.1037/0003-066X.52.10.1030

Stevens, S., Mills, R., & Kuchel, L. (2019). Teaching communication in general science degrees: highly valued but missing the mark. *Assessment & Evaluation in Higher Education*, 44(8), 1163-1176. DOI:10.1080/02602938.2019.1578861

Strasser, B., Baudry, J., Mahr, D., Sanchez, G., & Tancoigne, E. (2019). "Citizen science"? Rethinking science and public participation. *Science & Technology Studies*, 32(2), 52-76. DOI:10.23987/sts.60425

Strijbos, J., Engels, N., & Struyven, K. (2015). Criteria and standards of generic competences at bachelor degree level: A review study. *Educational Research Review,* 14, 18–32. DOI:10.1016/j.edurev.2015.01.001

Stupple, E., Maratos, F., Elander, J., Hunt, T., Cheung, K., & Aubeeluck, A. (2017). Development of the Critical Thinking Toolkit (CriTT): A measure of student attitudes and beliefs about critical thinking. *Thinking Skills and Creativity,* 23, 91-100. DOI:10.1016/j.tsc.2016.11.007

Styers, M., Van Zandt, P., & Hayden, K. (2018). Active learning in flipped life science courses promotes development of critical thinking skills. *CBE-Life Sciences Education*, 17(3), ar39, 1-13. DOI:10.1187/cbe.16-11-0332

Svihla, V., Wester, M., & Linn, M. (2018). Distributed practice in classroom inquiry science learning. *Learning: Research and Practice*, 4(2), 180-202. DOI:10.1080/23735082.2017.1371321

Szafranski, S., Wos-Oxley, M., Vilchez-Vargas, R., Jáuregui, R., Plumeier, I., Klawonn, F., Tomasch J., Meisinger, C., Kühnisch, J., Sztajer H., & Pieper, D. (2015). High-resolution taxonomic profiling of the subgingival microbiome for biomarker discovery and periodontitis diagnosis. *Applied and Environmental Microbiology*, 81(3), 1047-1058. DOI:10.1128/AEM.03534-14.

Taber, K. (2016). Learning generic skills through chemistry education. *Chemistry Education Research and Practice*, 17(2), 225-228. DOI:10.1039/c6rp90003h

Talanquer, V. (2014). DBER and STEM education reform: Are we up to the challenge? *Journal of Research in Science Teaching,* 51(6), 809-819. DOI:10.1002/tea.21162

Tan, K., & Gilbert, J. (2018). Changing practice: The impact of research. In J. Yeo, T. Teo & K. Tang (Eds.), *Science education research and practice in Asia-Pacific and beyond* (pp. 135-150). Singapore: Springer. DOI:10.1007/978-981-10-5149-4_10

Tapper, J. (2004). Student perceptions of how critical thinking is embedded in a degree program. *Higher Education Research and Development,* 23(2), 199-222. DOI:10.1080/0729436042000206663

Tarricone, P. (2011). *The taxonomy of metacognition*. Hove, England: Psychology Press.

Teasdale, R., Viskupic, K., Bartley, J., McConnell, D., Manduca, C., Bruckner, M., Farthing, D., & Iverson, E. (2017). A multidimensional assessment of reformed teaching practice in geoscience classrooms. *Geosphere*, 13(2), 608-627. DOI:10.1130/GES01479.1

Terenzini, P., Springer, L., Pascarella, E., & Nora, A. (1995). Influences affecting the development of students' critical thinking skills. *Research in Higher Education,* 36, 23-39. DOI:10.1007/BF02207765

The School Reform Longitudinal Study Research Team (2001). *School Reform Longitudinal Study: Classroom observation scoring manual.* St Lucia, QLD: Graduate School of Education, University of Queensland. Retrieved from https://core.ac.uk/download/pdf/15073826.pdf

Thomas, K., & Lok, B. (2015). Teaching critical thinking: An operational framework. In M. Davies & R. Barnett (Eds.), *The Palgrave handbook of critical thinking in higher education* (pp. 93-105). New York: Palgrave Macmillan.

Thompson, E. (1967). Time, work-discipline, and industrial capitalism. *Past & present*, 38, 56-97. Retrieved from https://www.jstor.org/stable/649749

Thompson, J. (2009). *To question or not to question: The effects of two teaching approaches on students' thinking dispositions, critical thinking skills, and course grades in a critical thinking course* (Doctoral dissertation, Capella University, Minneapolis, Minnesota). Retrieved from ProQuest Dissertations & Theses Global. (UMI 3355859).

Tiruneh, D., Gu, X., De Cock, M., & Elen, J. (2018). Systematic design of domain-specific instruction on near and far transfer of critical thinking skills. *International Journal of Educational Research,* 87(1), 1-11. DOI:10.1016/j.ijer.2017.10.005

Tiruneh, D., Verburgh, A., & Elen, J. (2014). Effectiveness of critical thinking instruction in higher education: A systematic review of intervention studies. *Higher Education Studies,* 4(1), 1–17. DOI:10.5539/hes.v4n1p1

Tiruneh, D., Weldeslassie, A., Kassa, A., Tefera, Z., De Cock, M., & Elen, J. (2016). Systematic design of a learning environment for domain-specific and domain-general critical thinking skills. *Educational technology research and development*, 64(3), 481-505. DOI:10.1007/s11423-015-9417-2

Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge, UK: Cambridge University Press.

Tremblay, K., Lalancette, D., & Roseveare, D. (2012). *AHELO feasibility study report - Volume 1: Design and implementation.* Paris, France: Organisation for Economic Co-operation and Development. Retrieved from http://hdl.voced.edu.au/10707/241317.

Trna, J., & Trnova, E. (2015). The current paradigms of science education and their expected impact on curriculum. *Procedia Social and Behavioral Sciences*, 197, 271-277. DOI:10.1016/j.sbspro.2015.07.135

Trujillo, C., & Long, T. (2018). Document co-citation analysis to enhance transdisciplinary research. *Science Advances*, 4(1), e1701130. DOI:10.1126/sciadv.1701130

Turner, R., Sutton, C., Muneer, R., Gray, C., Schaefer, N., & Swain, J. (2018). Exploring the potential of using undergraduates' knowledge, skills and experience in research methods as a proxy for capturing learning gain. *Higher Education Pedagogies*, 3(1).110-136. DOI:10.1080/23752696.2018.1449127

Tyser, R., & Cerbin, W. (1991). Critical thinking exercises for introductory biology courses. *Bioscience,* 41(1), 41-46. DOI:10.2307/1311540

Ulusoy, F., & Onen, A. (2014). *Developing the context-based Chemistry Motivation Scale: Validity and reliability*. *Journal of Baltic Science Education*, 13(6), 809-820. Retrieved from http://www.scientiasocialis.lt/jbse/files/pdf/vol13/809-820.Onen_JBSE_Vol.13_No.6.pdf

Ünsar, A., & Engin, E. (2013). A case study to determine critical thinking skills of university students. *Procedia Social and Behavioral Sciences*, 75, 563-569. DOI:10.1016/j.sbspro.2013.04.061

Vainikainen, M., Hautamäki, J., Hotulainen, R., & Kupiainen, S. (2015). General and specific thinking skills and schooling: Preparing the mind to new learning. *Thinking Skills and Creativity,* 18, 53-64. DOI:10.1016/j.tsc.2015.04.006

Vandermensbrugghe, J. (2004). The unbearable vagueness of critical thinking in the context of the Anglo-Saxonisation of education. *International Education Journal*, 5(3), 417-422.

van der Zanden, P., Denessen, E., Cillessen, A., & Meijer, P. (2018a). Domains and predictors of first-year student success: A systematic review. *Educational Research Review*, 23, 57-77. DOI:10.1016/j.edurev.2018.01.001

van der Zanden, P., Denessen, E., Cillessen, A., & Meijer, P. (2018b). Patterns of success: first-year student success in multiple domains. *Studies in Higher Education,* 44(11), 2081-2095. DOI:10.1080/03075079.2018.1493097

van Gelder, T. (2005). Teaching critical thinking: Some lessons from cognitive science. *College Teaching*, 53(1), 41–48. DOI:10.3200/CTCH.53.1.41-48

Van Gyn, G., & Ford, C. (2006). *Teaching for critical thinking.* Green guide No. 6. Ontario, Canada: Society for Teaching and Learning in Higher Education.

van Laar, E., van Deursen, A., van Dijk, J., & de Haan, J. (2017). The relation between 21st-century skills and digital skills: A systematic literature review. *Computers in human behavior*, 72, 577-588. DOI:10.1016/j.chb.2017.03.010

van Peppen, L., Verkoeijen, P., Heijltjes, A., Janssen, E., Koopmans, D., & van Gog, T. (2018). Effects of self-explaining on learning and transfer of critical thinking skills. *Frontiers in Education,* 3, 100. DOI:10.3389/feduc.2018.00100

Vincent, A., Decker, B., & Mumford, M. (2002). Divergent thinking, intelligence, and expertise: A test of alternative models. *Creativity Research Journal*, 14(2), 163-178. DOI:10.1207/S15326934CRJ1402_4

Viskupic, K., Ryker, K., Teasdale, R., Manduca, C., Iverson, E., Farthing, D., Bruckner, M., & McFadden, R. (2019). Classroom observations indicate the positive impacts of discipline-based professional development. *Journal for STEM Education Research*, 1-28. DOI:10.1007/s41979-019-00015-w

Vong, S., & Kaewurai, W. (2017). Instructional model development to enhance critical thinking and critical thinking teaching ability of trainee students at regional teaching training center in Takeo province, Cambodia. *Kasetsart Journal of Social Sciences*, 38(1), 88-95. DOI:10.1016/j.kjss.2016.05.002

Wagner, T., Baum, L., & Newbill, P. (2014). From rhetoric to real world: fostering higher order thinking through transdisciplinary collaboration. *Innovations in Education and Teaching International*, 51(6), 664–673. DOI:10.1080/14703297.2013.796726

Wall, T. (2015). The transferability of higher order cognitive skills. *Procedia Social and Behavioral Sciences*, 174, 233-238. DOI:10.1016/j.sbspro.2015.01.652

Walsh, C., & Seldomridge, L. (2006). Measuring critical thinking: One step forward, one step back. *Nurse Educator*, 31(4), 159-162. DOI:10.1097/00006223-200607000-00008

Wasserstein, R., Schirm, A., & Lazar, N. (2019). Moving to a world beyond "p < 0.05." *The American Statistician,* 73(sup1), 1–19. DOI:10.1080/00031305.2019.1583913

Watson, G., & Glaser, E. (1980). *Watson-Glaser Critical Thinking Appraisal.* San Antonio, TX: Psychological Corporation.

Watts, E., Levit, G., & Hoßfeld, U. (2016). Science standards: the foundation of evolution education in the United States. *Perspectives in Science*, 10, 59-65. DOI:10.1016/j.pisc.2016.08.004

Wechsler, S., Saiz, C., Rivas, S., Vendramini, C., Almeida, L., Mundim, M., & Franco, A. (2018). Creative and critical thinking: Independent or overlapping components? *Thinking Skills and Creativity,* 27, 114-122. DOI:10.1016/j.tsc.2017.12.003

Weinstein, M. (1995). Critical thinking: Expanding the paradigm. *Inquiry: Critical Thinking Across the Disciplines*, 15(1), 23-39. DOI:10.5840/inquiryctnews199515135

Weinstock, M., Kienhues, D., Feucht, F., & Ryan, M. (2017). Informed reflexivity: Enacting epistemic virtue. *Educational Psychologist,* 52(4), 284–298. DOI:10.1080/00461520.2017.1349662

Wickson, F., Carew, A., & Russell, A. (2006). Transdisciplinary research: characteristics, quandaries and quality. *Futures*, 38(9), 1046–1059. DOI:10.1016/j.futures.2006.02.011

Wiggins, B., Eddy, S., Grunspan, D., & Crowe, A. (2017). The ICAP active learning framework predicts the learning gains observed in intensely active classroom experiences. *AERA Open*, 3(2), 1-14. DOI:10.1177/2332858417708567

Wiksten, S. (2018). *Teacher training in Finland: A case study* (Doctoral dissertation, University of California, Los Angeles, California). Retrieved from ProQuest Dissertations & Theses Global. (UMI 10828213).

Wiliam, D. (2010). The role of formative assessment in effective learning environments. In H. Dumont, D. Istance & F. Benavides (Eds.), *The nature of learning: Using research to inspire practice* (pp. 135-155). Paris: Organisation for Economic Co-operation and Development Publishing.

Wiliam, D., & Black, P. (1996). Meanings and consequences: A basis for distinguishing formative and summative functions of assessment. *British Education Research Journal*, 22(5), 537-548. DOI:10.1080/0141192960220502

Williams, C., Walter, E., Henderson, C., & Beach, A. (2015). Describing undergraduate STEM teaching practices: A comparison of instructor self-report instruments. *International Journal of STEM Education,* 2(1), 18. DOI:10.1186/s40594-015-0031-y

Williams, M., & Burden, R. (1997). *Psychology for language teachers: A social constructivist approach*. Cambridge: Cambridge University Press.

Willingham, D. (2002). How we learn. Ask the cognitive scientist: Allocating student study time. Massed practice versus distributed practice. *American Educator,* 26, 37–39. Retrieved from https://www.aft.org/periodical/american-educator/summer-2002/ask-cognitive-scientist

Willingham, D. (2007). Critical thinking: Why is it so hard to teach? *American Educator,* 3, 8–19. DOI:10.3200/AEPR.109.4.21-32

Willingham, D. (2008). *Education for the twenty-first century: Balancing content knowledge with skills*. Retrieved from http://blogs.britannica.com/2008/12/schooling-for-the-twenty-first-century-balancing-content-knowledge-with-skills

Wilson, K. (2016). Critical reading, critical thinking: Delicate scaffolding in English for Academic Purposes (EAP). *Thinking Skills and Creativity,* 22, 256-265. DOI:10.1016/j.tsc.2016.10.002

Windle, M., Haardörfer, R., Getachew, B., Shah, J., Payne, J., Pillai, D., & Berg, C. (2018). A multivariate analysis of adverse childhood experiences and health behaviors and outcomes among college students. *Journal of American College Health*, 66(4), 246-251. DOI:10.1080/07448481.2018.1431892

Wittgenstein, L. (1958). *Philosophical investigations* (2nd ed.). Oxford: Basil Blackwell.

Wooten, M. (2019). A critical materialist entry into the comforts and dangers in defining academics' of science teaching and learning research communities. *Cultural Studies of Science Education*, 14(1), 231-247. DOI:10.1007/s11422-018-9871-3

Wright, C., Eddy, S., Wenderoth, M., Abshire, E., Blankenbiller, M., & Brownell, S. (2016). Cognitive difficulty and format of exams predicts gender and socioeconomic gaps in exam performance of students in introductory biology courses. *CBE-Life Sciences Education*, 15(2), ar23, 1-16. DOI:10.1187/cbe.15-12-0246

Wright, M., Bergom, I., & Bartholomew, T. (2019). Decreased class size, increased active learning? Intended and enacted teaching strategies in smaller classes. *Active Learning in Higher Education*, 20(1), 51-62. DOI:10.1177/1469787417735607

Wright, M., Goldwasser, M., Jacobson, W., & Dakes, C. (2017). Assessment from an educational development perspective. *To Improve the Academy*, 36(1), 39-49. DOI:10.1002/tia2.20051

Yanchar, S., Slife, B., & Warne, R. (2008). Critical thinking as disciplinary practice. *Review of General Psychology*, 12(3), 265. DOI:10.1037/1089-2680.12.3.265

Yang, F., Huang, R., & Tsai, I. (2016). The effects of epistemic beliefs in science and gender difference on university students' science-text reading: an eye-tracking study. *International Journal of Science and Mathematics Education*, 14, 473-498. DOI:10.1007/s10763-014-9578-1

Yang, Y., Newby, T., & Bill, R. (2008). Facilitating interactions through structured web-based bulletin boards: A quasi-experimental study on promoting learners' critical thinking skills. *Computers & Education*, 50, 1572–1585. DOI:10.1016/j.compedu.2007.04.006

Yazilitas, D., Svensson, J., de Vries, G., & Saharso, S. (2013). Gendered study choice: a literature review. A review of theory and research into the unequal representation of male and female students in mathematics, science, and technology. *Educational research and evaluation*, 19(6), 525-545. DOI:10.1080/13803611.2013.803931

Yetter, G., & Capaccioli, K. (2010). Differences in responses to Web and paper surveys among school professionals. *Behavior Research Methods*, 42(1), 266-272. DOI:10.3758/BRM.42.1.266

Yücel, B. (2008). An *evaluation of needs, design, implementation, and outcomes of Development and Learning course enriched with critical thinking based instruction* (Unpublished doctoral dissertation, Middle East Technical University, Ankara, Turkey). Retrieved from https://etd.lib.metu.edu.tr/upload/12609763/index.pdf

Zeegers, Y., & Elliott, K. (2019). Who's asking the questions in classrooms? Exploring teacher practice and student engagement in generating engaging and intellectually challenging questions, *Pedagogies: An International Journal*, 14(1), 17-32. DOI:10.1080/1554480X.2018.1537186

Ziman, J. (1980). *Teaching and learning about science and society*. New York: Cambridge University Press.

ŽivkoviL·, S. (2016). A model of critical thinking as an important attribute for success in the 21st century. *Procedia Social and Behavioral Sciences*, 232, 102-108. DOI:10.1016/j.sbspro.2016.10.034

Zlatkin-Troitschanskaia, O., Shavelson, R., & Kuhn, C. (2015). The international state of research on measurement of competency in higher education. *Studies in Higher Education*, 40(3), 393–411. DOI:10.1080/03075079.2015.1004241.

Zohar, A., & Barzilai, S. (2013). A review of research on metacognition in science education: current and future directions. *Studies in Science Education*, 49(2), 121–169. DOI:10.1080/03057267.2013.847261

Zohar, A., & Schwartzer, N. (2005). Assessing teachers' pedagogical knowledge in the context of teaching higher-order thinking. *Journal of Science Education,* 27(13), 1595–1620. DOI:10.1080/09500690500186592