Developing a tool for OSCE writers and reviewers to aid the identification of station-level errors.

Dr Kathy Brotchie

MBBS, FRACGP, DRANZOG,

Grad Dip Rural GP (Aboriginal Health)

September 15, 2015

A Thesis submitted in fulfilment
of the requirements for the degree of

Masters of Clinical Education by Research

Flinders University
Faculty of Medicine, Nursing and Health Sciences

Table of Contents

List of Tables

List of Figures

Acknowledgements

First and foremost I must acknowledge the dedication, skill and generosity demonstrated by my supervisors, Associate Professors Linda Sweet (Flinders University) and Shane Bullock (Monash University). Without their support, clever ideas and hours of meetings I have no doubt this thesis would not exist. I am truly blessed in having learnt from both of my supervisors during this journey.

Associate Professor Linda Sweet deserves additional recognition for being there at the very beginning of this process, and guiding me from the start to a solution which did not require a change in direction at any stage of this project. Additionally, the Flinders' boot camp weeks provided valuable intensive skills development that I can highly recommend to future distance education HDR students. Thanks also to those who taught me on the ground at Flinders University; I am very grateful for your direction and interventions.

I thank my colleagues Dr Marion Shuttleworth and Mrs Caroline Rossetti for sharing a vision for quality reform in our clinical skills program in Gippsland, and for demonstrating that attention to detail does reward both faculty and student and community. Thanks also to Caroline and Marion for their support for me during the past few years. Also Associate Professor Ray Tedman and Dr Cathy Haigh, for their willingness to contribute input to my work and for proof reading support and encouragement.

Finally to my family, and especially my husband Joël, who has taken on many roles over the past few months as I shed them in my determination to finish a project about which I have never lost interest. Your proof reading skills, ideas for alternative wording, and assistance with many presentations, have made your support and talents an integral part of my research achievements. I could not have done this without your faith in me.

To all of those I have mentioned and the many others who have helped along the way,

Thank you.

Summary

The core outcome of medical education programs is competent medical practitioners. Assessment of competence throughout all the stages of a doctor's career is necessary to ensure patient safety and effective practice. A popular approach to the assessment of clinical skills is the Objective Structured Clinical Examination (OSCE), originating in Scotland in 1975 and now globally accepted as a reliable and valid method. Despite its widespread adoption, concerns exist about the cost of conducting OSCEs, the impact on students of the high stakes examination process, authenticity issues, use of checklists and implementation errors including poorly written station items. The aim of this research is on improving the quality of the assessment of competence using the OSCE format, by aiding station developers and reviewers to identify station-level errors. The guiding research question is:

> *What aspects of the OSCE item writing process are prone to errors that undermine the quality of this assessment format and how can these be overcome?*

This thesis provides an insight into the concept of errors in OSCE stations used for the assessment of clinical competence of medical practitioners. The use of flawed stations undermines the candidate's opportunity to perform to the best of his or her ability, and ultimately reduces confidence in the results of that assessment. Importantly, a flawed assessment may prevent a competent doctor from becoming licenced to practice, reducing the available medical workforce and patient access to healthcare. Equally, a flawed assessment may allow incompetent doctors to practice unsupervised on patients or promote a medical student to the clinical environment when they are not yet ready to learn in that setting, compounding the learning deficit.

This project has met the aim by creating a tool to aid OSCE writers and reviewers to identify and correct errors affecting the validity, reliability, feasibility and educational impact of the assessment of clinical competence. Using a design-based research approach, a tool to aid in the quality improvement of OSCE station writing has been developed. Using a three phase iterative process, this thesis outlines the steps and decisions through the development of the OWSAT – the OSCE Writers Assessment Tool. This process has included considerable iteration, peer review at national and international conferences, and application of the tool against a database of OSCE stations.

Whilst not yet validated, the OWSAT contains questions highlighting aspects of OSCE writing that require reflection by the OSCE station writer or reviewer in search of improved station performance and overall assessment quality. It is anticipated that OWSAT will assist OSCE station writers and reviewers to identify errors at the writing or reviewing stage of the assessment process, and consequently enhance the systems that use OSCE assessment approach to determine competence in the medical profession.

Declaration of authorship

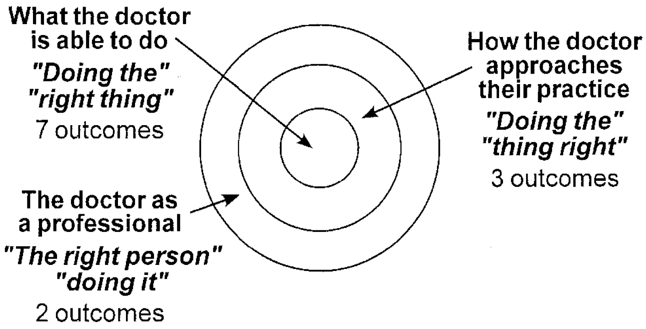
I certify that this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

Chapter 1 – Introduction

1.1 Assessing competence

Medical competence, as defined by Shumway and Harden (2003), includes not only what doctors can do, but also how they approach these skills using clinical reasoning within a legal and ethical framework. Assessment of competence throughout all the stages of a doctor's career is necessary to ensure an effective workforce. The therapeutic nature of the doctor-patient relationship relies on trust. For the patient, an implicit confidence results from an assumption that the medical professional being consulted will be competent. A model of competency was described in the setting of outcome-based learning by Harden and colleagues (1999) at the University of Dundee (see Figure 1-1).

Figure 1-1: The three-circle model for outcome-based education. (Harden et al., 1999, p. 547)


What the doctor is able to do
"Doing the" "right thing"
7 outcomes
How the doctor approaches their practice
"Doing the" "thing right"
3 outcomes
The doctor as a professional
"The right person" "doing it"
2 outcomes

Harden and colleagues' (1999) model emphasizes the active component and the breadth of a doctor's demonstration of competence. A guide to what needs to be assessed in an outcome-based approach to medical education, using the twelve outcomes for the three domains presented by Harden and colleagues (1999), is provided in Table 1-1.

Table 1-1: Expansion of the outcomes presented in Figure 1-1. (Harden et al., 1999, p. 550)

	Doing the right thing (7 outcomes)	Doing the thing right (3 outcomes)	The right person doing it (2 outcomes)
1	with competence in clinical skills	with understanding of basic, clinical and social sciences	with appreciation of the role of the doctor within the health service
2	with competence to perform practical procedures	with appropriate attitudes, ethical understanding and understanding of legal responsibilities	with aptitude for personal development
3	with competence to investigate a patient	with appropriate decision-making skills and clinical reasoning and judgement	
4	with competence to manage a patient		
5	with competence in health promotion and disease prevention		
6	with competence in skills of communication		
7	with competence to retrieve and handle information		

Competence in all aspects of clinical practice is desirable as doctors grow through their lifetime in their professional role (Shumway & Harden, 2003). Initially the measurement of competence is a responsibility of medical schools, marking the suitability of a student to progress from one year level to the next, or from the pre-clinical setting to the clinical years, and then finally to determine readiness to practice on graduation from university. After graduation, the onus of determining competence transfers through vocational training to specialty colleges, and ultimately becomes the domain of licensing and other regulatory bodies. A fundamental role of these organisations is to ensure that health professionals possess and retain the required knowledge, skills and behaviours to sustain trust and safety within any doctor-patient encounter.

Clinical competence is not only the domain of doctors; all health professionals are required to demonstrate knowledge, skills and attitudes relevant to their scope of practice (Liabsuetrakul et al., 2013). A term predominantly located within the nursing and allied health literature, 'scope of practice' is defined as performance of the duties and roles for which an individual has been educated, and adherence to relevant licensing and legal frameworks (Armstrong et al., 2005; Schuiling & Slager, 2000). Competence is context-dependent, with both environment and time playing a role in the transient nature of this construct. Whilst we cannot physically see 'competence', Dr Brian Hodges in his keynote address in Kuala Lumpur (2012) eloquently explained that we are nearly always able to answer the question, 'Would you send your mother to see this doctor? If not, why not? If yes, why yes?' Recognition of a competent doctor requires adequate opportunity for demonstration of professional performance in an authentic setting, incorporating observation of behaviours relating to ethical and legal frameworks.

Competence needs to be assessed, the assessment needs to be valid, and the assessment needs to include those aspects of practice that involve performance. A single observation may be incapable of identifying a doctor with a narrow skill base and a degree of luck in any given single encounter (Sauer et al., 2005). Direct observation of individuals performing these skills is an essential component in determining competence in a clinical setting as well as assisting through formative feedback in a non-clinical educational environment to develop expertise (Kogan et al., 2011). Assessment must therefore be fit for purpose, allowing observation of a doctor or student performing clinical tasks in the right context to enable the observer confidence in allowing them to practice in the clinical environment (Stiggins, 1997).

There are difficulties associated with assessing a doctor or medical student's level of competence. The verification of qualifications does not imply a consistent level of acquired knowledge or skills, or a currency of those skills (van Zanten et al., 2003). Assessment always involves making a judgement, and therefore is inherently subjective in nature (Schuwirth & van der Vleuten, 2004; Schuwirth & van der Vleuten, 2003). Fairness in an examination requires a 'level playing field' where all candidates are provided the same degree of task difficulty and the opportunity to perform the task to the best of their ability (Clarke et al., 2012). Fairness also requires assessors who are trained, impartial and who make judgements based on the observed task, not personal or personality-based criteria (Memon et al., 2010; Tversky & Kahneman, 1974). Ideally, the clinical skills assessment should permit competent doctors to display their skills and in the same setting identify the doctor who does not meet the required professional standards (Southgate et al., 2001b).

Many different clinical examination formats currently exist for assessing clinical skills competency, indicating that no single approach meets the needs of all health professional training and regulating bodies. Both the complexity of task and inherent imperfections in the existing structures led to the creation of many different assessment designs. Common assessments in medical education include the use of the long case, short cases, viva voce examinations, mini-clinical evaluation exercise (Mini-CEX), 360-degree assessments and the Objective Structured Clinical Examination (OSCE). A traditional method of oral assessment of clinical skills, the viva voce examination is described by Wakeford and colleagues as 'a general non-patient encounter between a candidate and one or more examiners' (1995, p. 931). A more contemporary option, the Mini-CEX, was originally designed as a formative assessment tool and is used to assess clinical skills where the observed encounter is with a real patient in a clinical setting and where feedback is given (Dijksterhuis et al., 2011). The Mini-CEX format is now widely used as both a formative and a summative tool applied to medical students in the clinical years. Borrowed from the corporate world, 360-degree assessments, also known as multisource feedback, involve gathering the opinions of peers, supervisors, non-medical colleagues and patients regarding performance in the workplace or clinical setting (Overeem et al., 2012). These assessment formats differ in the setting, presence of patient or actor playing the patient role, and the involvement of a passive or active examiner in the process. All assessment formats have raised concerns that have been well documented (Wakeford et al., 1995; Whitehead et al., 2015). The variety of formats of assessment allows individual institutions to meet their unique requirements for curriculum, reliability, validity and budgetary restrictions in assessing clinical competence. Whilst recognising the rich field of assessment options to research, this thesis is restricted to the popular, yet complex, approach for assessing competence in clinical skills assessment format known as OSCE.

Developed in the mid-seventies by Harden and Gleeson (1979), the OSCE is an examination of clinical skills where the candidates are required to perform clinical tasks in a simulated environment. Arora and Sevdalis (2008 p. 202), quoting McGaghie (1999), state that simulation is 'a person, device or set of conditions, which attempts to present evaluation problems authentically'. Just as Bracken and Rose (2011) argue that the 360-degree feedback process is an 'extremely complex process that requires dozens of nuanced decisions in its design and implementation' (2011, p. 184), so too, the OSCE is a complex undertaking requiring event-coordination-style preparations and multiple decision points during preparation and enactment. Decisions include creative choices on the content of the assessment of each individual task or station alongside clear communication of the tasks expected of all the participants in this practical assessment of clinical competence.

Health professions assessing undergraduate students, and later career professionals, have adopted the OSCE format globally. The use of OSCEs for the assessment of clinical skills has not been confined to the medical profession, with adoption of the format seen throughout the health professions. Dentistry, pharmacy, paramedics, midwifery and the nursing profession have all explored the use of the OSCE (Brosnan et al., 2006; Schoonheim-Klein et al., 2009; Sibbald & Regehr, 2003). There is a rich body of literature available to enable robust analysis of the advantages, disadvantages, and suitability of this assessment structure in health professions' education.

In an OSCE, each clinical task, or set of tasks, prescribed for the candidate in a particular location, to be performed within a set time limit, can collectively be considered as a station. Candidates progress from one station to the next in a circuit, performing clinical skills in a simulated environment, observed by a different examiner or examiners (Harden, 1988). Stations can vary in length, but the task or tasks should be realistic and achievable within the designated time frame. OSCE stations of five- or eight-minutes duration are common timings used by institutions globally. This time limit is thought to enable sufficient time to perform a physical examination of a body system or a focused history of a particular presentation. For example, the Royal Australian College of General Practitioners (RACGP) examination consists of twelve stations, each of eight-minutes duration, that represent the time taken in standard short consultations in general practice, based on data collected through the BEACH (Bettering the Evaluation and Care of Health) study (Britt et al., 2012). In addition, two longer-duration stations of nineteen-minutes duration are used to replicate the average long consultation times in general practice (Britt et al., 2012). Each individual station has a set time between tasks for the candidate to move to the next station and read the new task instructions, whilst the examiners record their assessment using a pre-determined marking system. In this thesis the term 'OSCE station' will be used to indicate a task or set of structured tasks occurring in the one location, including the components of reading time, task performance and marking by the examiner.

Arguments against the use of OSCE abound in the medical education literature. Criticisms include the high cost of running the examinations, and problems relating to authenticity, reliability, feasibility and educational impact (Norcini et al., 2011; van der Vleuten & Schuwirth, 2005; van der Vleuten & Swanson, 1990). A lack of care or skill in the design of an OSCE station, and the overall examination, may result in any or all of the above problems. For example, it is easy to create a poor OSCE station, one that provides insufficient details to enable a reliable assessment, or one that lacks authenticity. Close attention to the content wording of station details can address these criticisms. Such criticism may be viewed as cautioning against using OSCE uncritically, or they may be considered a challenge to improve the consistency of the OSCE framework or delivery. This research is a response to the second interpretation, with the intent of improving the consistency of the OSCE framework and/or delivery.

Assessments need to be valid, reliable, feasible and acceptable to all stakeholders (Southgate et al., 2001b). Reporting on the United Kingdom's General Medical Council processes of validation and revalidation of medical professionals, Southgate and colleagues advised (2001a, p. 4):

> *'If assessment methods are to be fit for purpose, they must be valid (particularly in discriminating between acceptable and unacceptable practice), and reliable. They must also be feasible, acceptable to the public and the profession, and robust enough to withstand legal challenge.'*

'Validity' means that the assessment has the ability to measure the attribute or attributes that it is intended to measure (Kane, 1992; Messick, 1995). 'Reliability' is defined by Cook and Beckman as the 'reproducibility or consistency of scores from one assessment to another' (2006, p. 166e.112). Breakdown in the public trust in the medical profession following a series of well-publicised adverse events in the United Kingdom led to the introduction of a robust system of assessments, including direct observation of clinical skills, for the purposes of revalidation (Donaldson et al., 2000). Importantly, the General Medical Council also provides remediation of identified poorly performing medical professionals who require further assessment of fitness to practice (Epstein & Hundert, 2002; Norcini et al., 2011).

Mitchell and colleagues (2009) reported on the use of OSCE in nursing, with a literature review of the key research articles exploring best practice in assessment through OSCE. Concerns about 'inconsistencies with the reliability and validity of the OSCE' (Mitchell et al., 2009, p. 400) resulted in the recommendation that a large number of stations of shorter duration be incorporated in nursing examinations, rather than the use of fewer stations of longer duration.

Failing to adhere to best practice principles of assessment within an examination may result in flawed decisions relating to the purpose of the examination. For example, in high-stakes post-graduate assessments such as for fellowship of a medical college, divergence from the best practice principles of assessment may prevent a capable doctor from practicing independently, or enable a doctor to practice when not competent (Klass, 1994; Norcini & McKinley, 2007). Similarly, decisions relating to undergraduate examinations may adversely impact the ability of an individual to progress to the next level, or to graduate from the course (Swanson et al., 1999).

Issues stemming from a poor understanding of assessment and evaluation processes are described in the literature relating to teaching, and are usually attributed to the lack of educational preparation in discipline-specific degree programs (Quitter, 1999). Medical education has traditionally been provided by clinicians and scientists, few of whom have any education qualifications or a deep understanding of assessment delivery (Taylor et al., 2012). In the standard medical school curriculum, future clinicians are rarely exposed to training in medical education or assessment principles. The concept of 'see one, do one, teach one' has been the basis of over a century of scientifically-based medical education process since the Flexner report of 1910 (Halperin, 2011), although the exact origins of this approach to teaching skills is unknown. Appointment to faculty in many medical schools does not depend on the possession of any qualifications in education (Hu et al., 2013). The possession of a medical degree is considered sufficient for most academic positions, although few medical-school curricula include modules in pedagogy or assessment. It is not surprising therefore, that an understanding of quality assessment principles may be lacking in those faculty tasked with writing OSCE stations.

A primary concern with the use of OSCE assessments is the quality of the written stations (Norcini & McKinley, 2007). Creation of the scenarios and the wording of the instructions provided to the parties involved in the station, such as the candidate, the simulated patient, the examiner, and the exam set up coordinator, is usually performed by a group of clinicians with the relevant clinical expertise (Klass, 1994). However, these clinical experts may not have the requisite educational knowledge, skills and experience in either the development of the stations, or the best practice principles of assessment. To overcome this concern, faculty development for case writers has been recommended (Boursicot et al., 2011), and the benefits of collaboration in the creation of OSCE stations is acknowledged (Brown & Skinner, 2003; Cohen et al., 1997). Furthermore, the significant workload associated with the creative development component of clinical skills assessment is also recognised within the academic literature, along with the role of faculty development for assessment-content developers as a determinant of quality (Boursicot et al., 2010; Schuwirth & van der Vleuten, 2013; Wamsley et al., 2013). Ensuring that the faculty member assessing the clinical skills possesses these skills is also a consideration in assessment quality, introducing the role played by factors beyond the item development phase (Holmboe et al., 2011).

The ability to recognise through self-assessment processes that an OSCE station is flawed may not be an aptitude present in all station writers. Eva and Regehr have summarised the literature on self-assessment, concluding that 'self-assessment…is inherently flawed' (Eva & Regehr, 2011, p. 312). Kruger and Dunning (1999) argue that an individual needs to possess the underlying knowledge to be able to self-identify errors in one's own work. Using the perspective of understanding English grammar and identifying grammatical errors, they argue that the capacity to recognise deficits in one's own creative output requires underlying knowledge that an incompetent individual may not possess (Kruger & Dunning, 1999). This will have an effect on self-monitoring ability. Eva and Regehr (2007) postulate a different competence, based on the skill of situational awareness, whereby the individual reflects on practice, as manifested by behaviours such as slowing down at the edges of competence, or knowing when to look things up. Eva and Regehr's papers (2007, 2011) provide possible explanations concerning how OSCE stations with major flaws in design or communication survive through to implementation, due to flawed self-assessment processes from incompetent or unskilled assessors. This work is anticipated to assist the creators of OSCE stations to gain requisite knowledge about and reflect on what is required in writing OSCE stations, thereby improving the item quality, using a process that will also be beneficial to OSCE reviewers.

Given that multiple factors contribute to the quality of assessment items, OSCE station writers should be acquainted with the core principles related to assessment fairness and defensibility (Schuwirth et al., 2002). Discussions around the quality of the OSCE may encompass the concepts of validity, reliability, reproducibility, utility, fidelity, generalizability, authenticity and practicality. Various authors of OSCE-related research publications have used selections of these terms, including Adreatta and colleagues exploring validity in the assessment of Obstetrics and Gynaecology skills (2011), and Norman, who published an article explaining reliability (2014). Swanson and Norcini discussed reproducibility (1989) as did Downing (2004) who linked the concept to the scientific method where 'experiments must be reproducible in order to be properly interpreted or taken seriously' (p. 1006). Utility was added to the discussion regarding quality in assessment by van der Vleuten and Schuwirth (2005), but was also the subject of papers by llgen and colleagues (2013) and Gorsira (2009). The foundations of this form of assessment require original thought and creativity in the setting of the convergence of clinical theory and assessment principles (Vargas et al., 2007). Station writers who lack understanding of the relevant clinical details or the technical aspects of creating a simulation scenario may produce an unfair or unfeasible assessment (Tavakol & Dennick, 2011; Townsend et al., 2010).

Reliability and validity are vulnerable to errors in the station-writing process, and improving the written stations is one component of improving the assessment process. Whilst acknowledging the popularity of OSCEs as an assessment format (Schuwirth & van der Vleuten, 2003) used globally by many health professions, the spotlight here is on assessment in medical training. This will include undergraduate medical student assessments, vocational assessments for different specialist groups, and international medical graduate examinations. Developing a structured approach to the identification of errors is the key outcome of this thesis. Through the station-writing process we can influence one component of the assessment process with the aim of improving the overall quality of the assessment of clinical competence.

1.2 Statement of the Problem

The prospect of spending years exploring a topic through the academic processes of higher degree research suggests a likely personal motivation for the choice of research topic, beyond curiosity. Matching core values underpins most successful long-term relationships, and understanding one's own value system forms one dimension of the personal journey undertaken through the research process. My values include fairness as a key driver in my expectations for quality in assessment. The journey undertaken just prior to the commencement of this thesis, that of successfully negotiating the Canadian medical licensure process, led to a decision to explore best practice in OSCE processes from the perspective of a candidate. My other roles as an examiner at a Canadian medical school, as well as for the Royal Australian College of General Practitioners, gave me a different perspective as I watched a series of candidates undertake the prescribed task based on their interpretation of the script. Both of these experiences in the OSCE format led to expectations of how stations should provide a fair opportunity for a candidate to perform requested tasks. The reading time should enable candidates to effectively plan their time and how they will approach the station without being undermined by unnecessary interruptions. Identification of many stations that fail to include on the candidate's instructions sufficient information for the planning of the task is just one of many issues identified personally in a world where no standard exists for the content or language employed in OSCE writing.

Serious decisions are determined by the results of high-stakes examinations of the clinical competence of medical professionals every year. Confidence in the quality of these decisions should not be presumed, but based on a firm evidence base. The effect size of any one error is impossible to quantify in this complex interaction between skill, judge, patient, candidate and environment. Reducing the influence of errors in our assessments may not even change the examination results; however, assuming errors do not occur in the OSCE setting has risks for patients and for medical professionals.

1.3 Aim of the Research

The aim of this research is on improving the quality of the assessment of clinical competence using the Objective Structured Clinical Examination format, by aiding station developers and reviewers to identify station-level errors. The guiding research question is:

> *What aspects of the OSCE item writing process are prone to errors that undermine the quality of this assessment format and how can these be overcome?*

Explicitly, this project pursues a way to improve the quality of OSCE station writing. Development of a tool to aid OSCE station writers and reviewers to understand the decision-making around OSCE station development will be explored as a potential solution to concerns about quality within this assessment format.

1.4 Thesis Structure

In the seven chapters of this thesis the topic of errors in the written item content within the assessment framework known as OSCE will be investigated. These chapters contain the background and process involved in the development of a tool to assist OSCE writers and reviewers to identify possible flaws in OSCE stations. Undertaken during the period between March 2010 and June 2015, both the tool development phase and the review of the literature published during and prior to these dates involved an iterative process.

Chapter 1 has introduced the notion of competence assessment, the importance of valid assessment, and the popular format of the OSCE for assessment of competence in the health professions.

Chapter 2 presents a review of the literature relating to OSCE; the history, benefits and criticisms of this format and the psychometric principles used in the evaluation of clinical assessment. Errors and the use of systems to detect them, the concept of assessment literacy and the importance of language in the writing and development of OSCE are explored through an analysis of the available literature. There is a gap in the literature relating to explicit acknowledgement of the issue of errors in OSCE station writing; hence methods to address this problem are examined in this chapter.

Chapter 3 focuses on methodology, and discusses the rationale behind the choice of design-based research to address the problem of how to identify errors in OSCE stations. The decision to develop a tool using the iterative design-based research approach is introduced in this chapter. Steps undertaken to obtain ethics approval and the three phases of the research project are outlined in this chapter.

Phase one is described in Chapter 4. This chapter explores the background and context relating to the identification of existing errors in actual OSCE stations and the exploration of literature relating to this topic. Information obtained relating to station-level flaws through the post-OSCE debrief process is presented in this chapter. Phase one concludes with the creation of a list of OSCE errors and questions to consider during station development.

Chapter 5 contains the details about phase two of the iterative approach to the creation of the OSCE Writers Station Analysis Tool (OWSAT) and the process of peer review of the early iterations of this instrument. The quality improvement of the tool through presentations at international conferences on the topic of OSCE errors and the refinement of the tool are discussed in this chapter.

Chapter 6 explores the third and final phase of this design-based research project. Use of a database of existing OSCE stations gathered from local and international medical education institutions to field test the tool is described in this chapter. Construction of an electronic version of the tool and subsequent iterative modifications are detailed in this phase.

The final chapter summarises the work undertaken in the creation of the tool to address OSCE station-level errors and the justification of the approach taken in this research. Limitations of the project and how these might be addressed in the future are detailed in Chapter 7. Future research directions using the tool to assist OSCE writers and reviewers to identify flawed OSCE stations are suggested in this chapter.

Chapter 2 – Literature Review

2.1 Introduction

This chapter outlines the academic literature pertaining to assessment in the domain of clinical competence using the OSCE, and presents key articles outlining its strengths and weaknesses as an assessment tool. A review of the body of work amassed since the mid-1970s was undertaken, exploring performance-based assessments of clinical competence, use of simulation in assessment, theory relating to errors and quality control processes encountered in medical education assessment. Whilst the focus of this thesis is on medical vocational training, and undergraduate medical education, the literature examined includes other health professions, education, psychology and business fields. This body of literature contains important works relating to performance assessment and understanding of psychometric principles.

This literature review involved an iterative process, which cycled between structured reviews of traditional research databases and the trawling of web-based browsers such as 'Google'. As a novice medical educator seeking to locate any information to assist with the process of writing OSCE stations, this search for best practices and relevant information aided in understanding the utility of key words and information technology resources. Aiming to improve understanding of all aspects of OSCE development this review searched specifically for publications providing solutions relating to quality concerns in OSCE station design. Beginning with my coursework degree, key articles about OSCEs recommended by experts in the field of medical education research were collated. Observing the choice of key articles by experts presenting assessment workshops at international conferences aided further in providing the evidence base for the use and conduct of the OSCE format of examinations.

The use of database search tools such as OvidSP®, which includes both MEDLINE® and ERIC℠ (Educational Research Information Centre), provided a more systematic approach to the review of literature. These two repositories cover medical and educational literature respectively. In addition, both basic and advanced searches were undertaken through the university's online library resources. Search terms used within the OvidSP® database included OSCE, Objective Structured Clinical Examination*, error* and quality.

Initial searches with Google and university library search facilities required exclusion of data relating to the other 'OSCE' – the Organization for Security and Co-operation in Europe. This term gained broad recognition following the loss of Malaysian Airlines flight M17 over the Ukraine in July 2014. Repeated searches immediately after July 2014 did not require this exclusion. Using the search term OSCE or Objective Structured Clinical Examination revealed over 6000 potentially relevant articles. Combining with search terms such as 'error*' or 'station-level' produced significantly fewer results and combining all three within the OvidSP® database produced no results.

The use of both the Monash University library 'SEARCH' option and the Flinders University library 'FindIt@Flinders' provided a unique opportunity to access a breadth of literature through the use of remote library access to on-line facilities. Both provide access to all the resources the universities have in their collections, as well as full text on-line access to journals to which the universities have subscribed. There were differences in availability between the two university libraries. Similar content existed in both, but the search engines gave different results to the same search terms and there were perceived different priorities with respect to full access subscriptions.

A search within the Informa Healthcare database for 'OSCE' and 'station errors' yielded 149 results, many of which were duplicates of those already located. Most of the articles contained within this search were related to the implementation of the OSCE as an assessment format for observable technical and non-technical clinical skills over the past four decades. Some of the literature related to concerns about the analysis of results from OSCEs and performance-based assessment. Very few contained any detail about the station-level aspects of the examination and their contribution to the validity and or reliability of the OSCE.

These searches, including the original OvidSP® search, and even a Google search were repeated multiple times during this project with additional journal articles and several books being added every year. Scrutiny of the reference list of book chapters and journal articles contributed several additional articles. The reference lists obtained at workshops I attended on assessment in medical education, such as the ESMEA (Essential Skills in Medical Education Assessment) course and the Fundamentals of Assessment in Medical Education Course (Course, 2012), contributed further additions to the literature bank and my knowledge of assessment principles and procedures.

Finally, a Google search using 'OSCE quality improvement' with no language restriction identified a single article written in German, but with an English translation of the abstract (Schultz et al., 2008). This did not pose a problem for me, as I am able to read and speak German. As this particular article provided the only publication which covered the research question relating to tools for assisting OSCE item writers to improve the quality of their work, the full document was carefully translated into English and the authors contacted in person to explore more thoroughly the details surrounding this published work. The failure to identify within the available English language literature any serious considerations of the nature and impact of station-level flaws, apart from this German published work, provided challenges for the overall search process. Despite an extensive and repeated review process, with the exception of the single article from Heidelberg University, Germany, no other work closely matched the intent of my project to improve station-level OSCE quality.

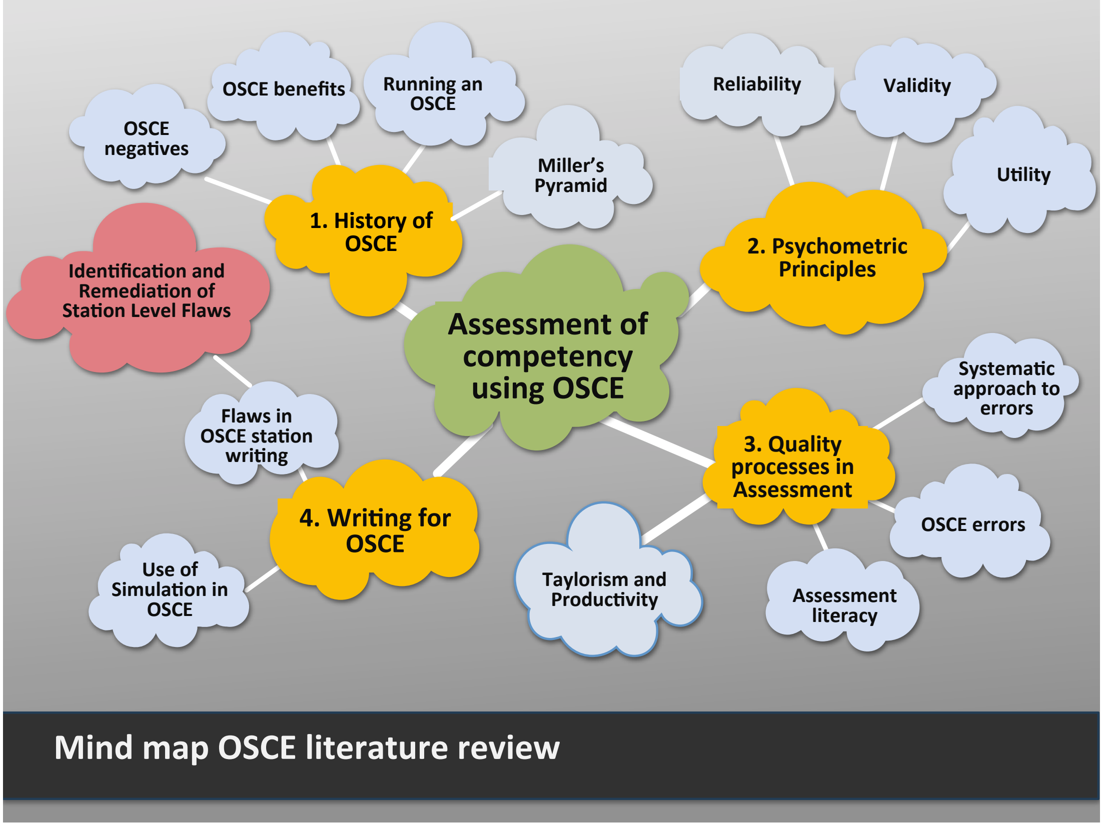
The use of limitations in this review included restricting performance-based assessment to the OSCE format, inclusion of articles predominantly relating to medical students and vocational assessments in medicine and restricting the searches to English-language publications. However, the paucity of articles specifically addressing OSCE station-level writing made limitations temporary. Other restrictions were not helpful as the potential for locating relevant information regarding the effect of station-level errors occasionally emerged from within the substance of an article on the topic of OSCE, irrespective of the title or keywords in the publication. Further limitations including restricting the search to the last five or ten years would have been unproductive as much of the relevant literature regarding station design was published in the late 1970s and early 1980s. This represented the decade immediately after the revelation to the medical education world of the innovation known as the OSCE. Exploration of subsequent decades covered the emergence of criticism regarding the examination format, the publication of systematic reviews on the topic prior to implementation in other health professional domains, as well as other relevant works regarding the data analysis and standard setting approaches to the use of OSCEs.

There are many books now published on the topic of the OSCE. The majority of these contain information for students, and are for candidates planning or preparing to sit this format of assessment. The exception to this involves two books, both entitled 'OSCE', published within the past five years. The first is a book by Brian Hodges (2009), exploring the history of OSCEs, and the other book published in 2013 by Zabar and colleagues on how to run an OSCE. Relevant literature to assist faculty with this sometimes high-stakes, expensive and time-consuming endeavour remained elusive until this recent work emerged. Still failing to address the concept of reducing assessment validity through implementation of flawed stations, the publication by Zabar and colleagues (2013) provided the most comprehensive details yet on how to conduct this form of clinical assessment.

2.2 Discourse on OSCEs

Best described as a thematic review of the literature pertaining to OSCEs, this analysis examined materials from over 800 articles sourced from the popular, grey and refereed literature. Four themes emerged during the literature review process associated with the assessment of clinical competence. Beginning with the origin and implementation of the OSCE, the other themes include understanding the psychometric principles relating to assessment, exploring quality in assessment, and writing for structured clinical examinations. The structure of this literature review is presented in Figure 2-1.

Figure 2-1: Structure of OSCE literature review


OSCE benefits
Running an OSCE
Reliability
Validity
OSCE negatives
Miller's Pyramid
Utility
1. History of OSCE
2. Psychometric Principles
Identification and Remediation of Station Level Flaws
Assessment of competency using OSCE
Systematic approach to errors
Flaws in OSCE station writing
3. Quality processes in Assessment
4. Writing for OSCE
OSCE errors
Use of Simulation in OSCE
Taylorism and Productivity
Assessment literacy
Mind map OSCE literature review

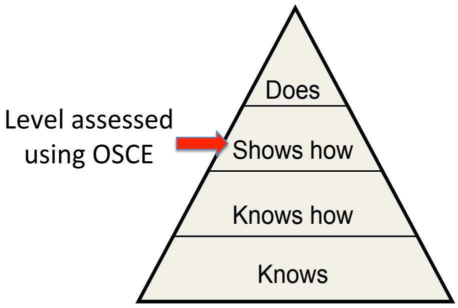
This thesis required an examination of works from all four decades for the following reasons: 1) The early works contain the best descriptions of OSCE stations, information which was often glossed over in subsequent decades leaving the reader with the assumption that the item writing process was flawless and unlikely to contribute to the decisions being made on the relevance, reliability or feasibility of format. 2) The works exploring applicability to other settings contained information about the administration aspects of the OSCE process and gave insight into whether and in what format the station details and quality assurance processes surrounding their construction existed as the assessment format transitioned into new domains. 3) The publications criticising the OSCE for lacking various qualities contained deductions potentially biased through station flaws. Elements requiring correction as described in these works may not have existed if attention to station design details improved. Providing insight into things that did not go well during an OSCE, these issues initiated the interest leading to this research project. And finally 4) the meta-analyses of OSCE and the reminiscing type literature has the potential to provide further understanding of how well the comprehension of station-level flaws impacting on standardized performance or validity of assessment has passed into the awareness of those who run examinations using OSCE formats.

The past four decades traversed changes in the focus of research relating to the OSCE. Hodges' doctoral thesis (2007) explored the OSCE using a Foucauldian Genealogical approach, identifying three defined discourses in the history of OSCE since its inception in 1975. These themes began with the concept of observing performance in assessment which he labelled 'Millers' pyramid and performance' discourse (Hodges, 2007 p 62). Subsequently, the 'Cronbach's alpha and psychometric' discourse (Hodges, 2007, p. 113) explored the quest for standardisation to achieve improved reliability. The final influence on the OSCE research and institutional focus was termed the 'Taylorism and production discourse' (Hodges, 2007, p. 158). Despite 100 years since the publication of Taylor's seminal work, 'The Principles of Scientific Management' (Taylor, 1911), the influence of this approach to business, education and assessment is still relevant and underpinned the third discourse in Hodges' socio-historical review of the OSCE. Hodges' perspective on the evolution of literature relating to the OSCE aligns with the mind-map image of this thesis' literature review. Each discourse on the OSCE represents a different but relevant component that must be considered with respect to errors in OSCE stations. The use of the OSCE to examine the most appropriate level of Miller's pyramid is fundamental to station design and content considerations. Quality in OSCE was determined predominantly using the psychometric discourse, while application of elements of Taylor's discourse has the potential to aid in the solution to the dilemma of OSCE station-level flaws.

2.2.1 Miller’s Pyramid and the Dreyfus Five-Stage Model of Skills Acquisition

Fundamental to understanding the best application of the OSCE in assessing competence in medical professionals, Miller’s Pyramid separates the ‘knows how’ from the ‘shows how’ in the diagrammatical representation of assessment hierarchy (Miller, 1990). Miller’s desire to see knowledge and skills demonstrated rather than merely discussed (Miller, 1990), and the psychometricians’ quest for improved standardisation, are easily recognisable dialogues that have shaped the literature and practices of the assessment of clinical competence (Schneid et al., 2014) . Miller described four levels of assessment relating to medical education, and displayed these in the form of a pyramid. These levels were
1) demonstrating knowledge, 2) knowing what to do, 3) showing what to do, or 4) actually doing, with respect to patient encounters (Miller, 1990). The ‘shows how’ level is accepted as being assessed by the OSCE (Khan et al., 2013b). This pyramid of assessment is shown in Figure 2.2.

Figure 2-2: Miller's Pyramid (Miller, 1990, p. S63)



Miller GE. The assessment of clinical skills/ competence/performance. Academic Medicine (Supplement) 1990; 65: S63-S7.

An alternative to Miller’s Pyramid is found in Dreyfus’ competency model as described by Pyrani and colleagues (2013), in their approach to early clinical skills at a Nepalese medical school. Dreyfus and Dreyfus (1980) explored the human path to expertise, interviewing pilots, chess players and language students in their search for an artificial intelligence application. The Dreyfus ‘Five-stage model’ identified a progression in skills development from novice to expert with changes in mental functions as skill level increased (Dreyfus & Dreyfus, 1980, p. 15). These stages are reproduced in Table 2-1. An interpretation of the terms used in the Dreyfus model as applied to education is shown in Table 2-2.

Table 2-1: Dreyfus Five-stage Model of the Mental Activities Involved in Directed Skills Acquisition (Dreyfus & Dreyfus, 1980, p. 15):

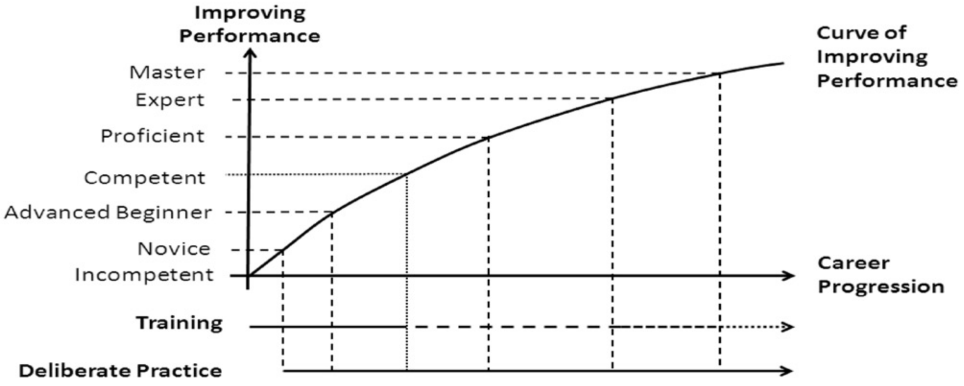
Skill Mental Function	Novice	Competent	Proficient	Expert	Master
Recollection	Non-situational	Situational	Situational	Situational	Situational
Recognition	Decomposed	Decomposed	Holistic	Holistic	Holistic
Decision	Analytical	Analytical	Analytical	Intuitive	Intuitive
Awareness	Monitoring	Monitoring	Monitoring	Monitoring	Absorbed

Table 2-2: Interpretation of Dreyfus Five-stage Model (Dreyfus & Dreyfus, 1980, p. 15)

Situational	Occurs with recognition of previous experience
Non-Situational	Is context free
Decomposed	Initially tasks are broken down into pieces
Holistic	Ability to see the whole task
Analytical	Calculations are being made to aid decision making
Intuitive	Decisions are made without the need for calculation
Monitoring	Mind has a monitoring role to produce and evaluate performance
Absorbed	Mind is freed from these duties completely absorbed in performance

Khan and Ramachandran (2013b) cite the Dreyfus model and reject the assumption that the OSCE is an assessment of competence. Khan and Ramachandran argue that the OSCE 'is a tool for the assessment of performance within simulated environments' (2013b, p. e1440). Whilst the Miller's Pyramid has had widespread adoption by researchers into health professional assessment, the Dreyfus model has more applicability to the recognition of different levels of learner, and has implications for the observed differences in OSCE writing and reviewing capability. A graphic representation of this journey from novice to expert, modified from Dreyfus and Dreyfus (1980) and Olle ten Cate and colleagues (2010) is presented in Figure 2-3 (Khan & Ramachandran, 2012).

Figure 2-3: Curve of improving performance adapted for healthcare (Khan & Ramachandran, 2012, p. 922).



The discourse of observing performance as a desired inclusion in assessment preceded the introduction of the OSCE and continues to be relevant today (Hodges, 2007). The second and third discourses relating to the history of OSCE in the research literature, namely the psychometric discourse and the Taylorism discourse, will be discussed within the results of the literature review.

2.3 The OSCE as an assessment format

2.3.1 History of the OSCE

First described in 1975 by Harden, Stevenson, Downie and Wilson in the British Medical Journal, the OSCE is an assessment of clinical skills where the content of the examination and the standards required of the students are determined prior to the examination. Initially, many published works provided some definition of OSCE using a variety of descriptive terms. In 2005, a definition of OSCE provided by Boursicot and Roberts gave a practical depiction of the OSCE as 'an assessment format in which the candidates rotate around a circuit of stations, at each of which specific tasks have to be performed, usually involving a clinical skill such as history taking or examining a patient'(p. 16). This approach allows a more visual understanding of the concept providing clear expectations for simulated patients, examiners and candidates. More recently in 2013, Fuller, Homer and Pell make no attempt to define the OSCE, but indicate it is a highly credible assessment tool used in examinations for clinical performance.

The OSCE format arose out of concerns about fairness and reliability within the established clinical assessment options. Throughout the 1970s, recognition of the problem of significant variation in marking by paired examiners using the clinical examination formats of the time, had led to calls for an alternative form of assessment (Fleming et al., 1974; Harden, 1979a). The model termed 'OSCE' introduced by Harden and colleagues (1975) in their landmark paper was innovative and claimed to avoid many of the flaws associated with the traditional observed clinical competency assessments. However, the authors highlighted two main concerns about the potential impact of the assessment design; that of having students compartmentalize patients due to the reduction of the whole patient examination into smaller tasks, and the amount of preparation required to run this extensive examination (Harden et al., 1975).

Despite these concerns, an early AMEE guide (Harden & Gleeson, 1979) provided details of this format enabling medical schools around the world to adopt this approach to assessment of competence. Many academic institutions began using the basic OSCE framework, adapted to their own budgets and needs. By 1985, Harden reported significant variation in the organization of the OSCE at different centres whilst retaining overall consistency in the process of students rotating through a number of stations, each of which is testing a different clinical skill, with the use of checklists and observing examiners for each procedural task station (Harden, 1985).

The adoption of the OSCE within the United Kingdom in 1995 as part of the assessment process for poorly performing doctors indicates the extent to which it had become an accepted method of determining competence in practical skills (Southgate et al., 2001a). Casey and colleagues (2009) identified the level to which the OSCE had become part of medical school assessment processes with the United States of America. Of the 126 US medical schools surveyed by the Liaison Committee for Medical Education, 97 were using the OSCE format in some context for assessment of clinical skills (Casey et al., 2009, p. 25). The move to global acceptance within the three decades following the introduction of the OSCE highlights the extent to which the evolution of the assessment of clinical competence has gained in importance using this format.

Research publications on the topic of OSCEs underwent significant changes in emphasis over the decades since the original articles by Harden and his colleagues. A typical evolution of any idea follows a similar process. During the decade following the depiction of the structured assessment format there were many published works detailing descriptions of how the OSCE was being implemented, with minor variations being used as case studies. Alongside this were detailed analyses of standard setting, statistical verification and publications demonstrating utility in other settings within the health professions. The onset of a series of works criticizing the mode of assessment were published in the 1990s alongside academic works exploring authentic workplace-based assessments using standardized marking tools. Finally more recently meta-analyses of various aspects have been published alongside the aforementioned books on how to run, sit or understand the place in assessment held by the OSCE as it has been portrayed in the research literature published in the past four decades.

2.3.2 The Advantages of the OSCE

Institutions undertaking OSCEs have autonomy over the many decisions required in the administration of this type of examination. Van der Vleuten and Swanson (1990, p. 59) argued that the OSCE is not an assessment format, but a 'flexible approach to test administration in which a variety of methods can be embedded to obtain an assessment of clinical skills'. Advantages of the OSCE include this flexibility, enabling adopters of this 'approach' to adapt the structure to the local needs in terms of content, timing, number of stations and qualifications/preparation of assessors. Decisions required within the structure of an OSCE station and examination include how the marking guide is set up, the role of simulated patients and examiners in terms of the task created, and processing and degree of analysis of the results. Ultimately the flexibility of the OSCE format enables an institution to arrive at an agreed assessment of clinical competence based on the purpose of the assessment, local preferences, and available resources.

2.3.3 The Disadvantages of the OSCE

Criticisms of the OSCE format were identified within the medical education literature, matching the reflection of Hagen and colleagues (1994) that no process of assessment lacked criticism. The list of pre-existing concerns relating to competency-based assessments provided fuel for some OSCE related criticisms (Anonymous, 1876; Fleming et al., 1974; Harden, 1979b). Competency-based assessments are complex, and have been described as looking for the end product of education, including the behaviours required to perform a task. The interaction of a candidate with a simulated patient whilst performing a procedural task (Albanese et al., 2010) such as an intra-muscular injection into a part-trainer attached to the patient's arm is an example of a competency that may be assessed through the use of OSCE. Criticisms of performance-based assessment, observation of the specific task or tasks rather than a higher order of assessment of professional behaviours during a procedure or performance of history or examination skills, are also frequently encountered in the medical education literature (Maatsch, 1981). Concerns relating to the assessment of performance or competencies include issues with validity, reliability, and being only 'capable of dealing with the superficial or trivial' (Gonczi, 1994, p. 43).

Concerns about content specificity have been commented upon by several authors (Swanson & Norcini, 1989; Townsend et al., 2010; van der Vleuten et al., 2010). Defined as the inability of a candidate to perform at the same level from one case or station to another, even where the content was similar, content specificity has also been described as case specificity (Swanson & Norcini, 1989, p. 158). The variability of candidate performance is a feature easily observed during OSCEs. A proposed solution uses large numbers of cases to diminish the effect of the variation on the reliability of results (Swanson & Norcini, 1989). This remedy is more easily achieved using the OSCE format than finding sufficient numbers of real patients for observed clinical performance assessment (Levine et al., 2012).

2.3.4 Running an OSCE

Many authors were critical of the expense associated with running OSCEs (Turner & Dankoski, 2008; Walsh et al., 2013; Walters et al., 2005). Varkey and colleagues (2008) dismissed criticisms of OSCEs related to running costs, claiming an implementation cost of $255 per fellow as reasonable; however they confessed that they had not included the 45 hours of faculty time required for developing their 8 OSCE stations in their cost analysis (Varkey et al., 2008). Walsh and colleagues recommended the use of cost-utility analyses given the already high cost of medical education (2013). Walsh and colleagues conclude that an intervention such as OSCE or simulation based assessment may be worthwhile if it prevents patient complaints or improves the quality of graduates from a medical program (Walsh et al., 2013).

Even back in 1979, Harden commented that some criticism of assessment may be due to the poor quality of the assessment that then fails to fulfill the intended function or was inappropriate for the assessment task required (Boursicot & Roberts, 2005; Harden & Gleeson, 1979). It can be argued that many of today's versions of the OSCE might contain both of these faults (Boursicot & Roberts, 2005). Examples of this include assessing students' knowledge, which may be better tested in a written examination, or other errors that impair the quality of the overall exam by undermining validity or reliability.

More publications criticizing the OSCE emerged in the 1990s (Hager et al., 1994; Reznick et al., 1993b; van der Vleuten et al., 1991). Many of these occurred within published academic works exploring authentic workplace-based assessments using standardized marking tools. Research protocols and traditions, including those for higher degrees, require researchers to criticise and find a gap in the available literature to enable positioning for their own research output. Publications on tools such as the Mini-CEX and other in-training assessment formats as well as literature on a programmatic approach to assessment emerged in the past decade as the quest for addressing the ‘does’ apex of Miller’s Pyramid (Miller, 1990, p. s63) rather than the ‘shows how’ segment covered by the OSCE (Bok et al., 2013; Schuwirth & van der Vleuten, 2011b). More recently, meta-analyses of various aspects have been published alongside the aforementioned books on how to run, sit or understand the place in assessment held by the OSCE as it has been portrayed in the research literature published in the past four decades (Khan et al., 2013b; Mitchell et al., 2009).

2.3.5 Use of simulation in OSCE

Simulation is a key element of most OSCE stations (van der Vleuten & Swanson, 1990). Simulation using people as actors and portraying clinical signs was first developed by Howard Barrows in 1963 when he employed an art class model to portray lower limb paralysis because he thought she ‘wouldn’t get upset’ whilst having her legs examined by medical students (Barrows, 1987, p. viii). Adamo (2003) praised the use of simulated or standardised patients in the OSCE format, with multiple patients behind doors of a corridor as being a realistic interpretation of a clinical setting. Klass (1994) reported the rise in the popularity of simulation in the setting of high-stakes assessments for national licensure examinations in the USA, whilst cautioning that the medical profession at the time was unfamiliar with the use of simulated patients and would need to develop faith in this tool for testing and teaching (Klass, 1994). The international experience of simulation using actors to portray roles in medical education and assessment was explored from the perspective of creating a bank of globally relevant cases (Sutnick et al., 1994).

Many assessment formats involve the use of simulation, whilst others, e.g. the Mini-CEX, rely on the traditional use of bedside teaching and real patients. The OSCE may involve real patients but is predominantly reliant on simulation and the recreation of the clinical environment in another location with actors or volunteers portraying the role of patient from a scripted medical case. A simulated patient or clinician may play the role of the examiner, and the location of the examiner may vary from sitting in the consultation room, or watching from another location via video recording (van Zanten et al., 2003). The use of simulation to reproduce a patient encounter in a non-clinical setting has enabled standardisation of this interaction for the purposes of education and assessment (Sutnick et al., 1994). Simulation also allows simultaneous examination of large numbers of people doing the same set of tasks (Boulet, 2008). Performances can then be compared, and candidates assessed for their level of clinical competence and benchmarked against one another (Boursicot et al., 2007). Multiple tasks can be performed within one encounter between an SP and candidate, requiring complex scripts to be constructed carefully in multi-circuit or venue OSCEs to ensure a fair and equitable assessment (Swanson & Norcini, 1989; van der Vleuten & Swanson, 1990).

2.3.6 Human resource requirements

As discussed previously in the Discourse on OSCE section, Hodges referred to the more recent literature exploring the human resources and administration requirements, including standardisation of processes using the Taylorism metaphor (2007). Emphasising aspects of the OSCE beyond the concept of assessing the 'doing' or analysing the psychometrics of examination results, Hodges' 'Taylorism' discourse identified the creation of new roles in the evolution of OSCE during the past decade. Simulated patient trainers and the creation of testing centres to run national examinations are reminiscent of the scientific approach to management espoused by Taylor (1911).

A key role in the OSCE process is that of the writer of individual stations. This role can be shared amongst faculty, but attempts to get consensus are easier with fewer doctors in the room, as differences in opinion relating to requisite standards are well documented (Kogan et al., 2011; Liao et al., 2010). Research conducted by Wilkinson and colleagues into the level of involvement by examiners in OSCE development suggested many examiners are not involved in this process (2003). Using a five point Likert scale, the study resulted in a median score of 2.0 where 5 was complete involvement and 1 was no involvement (Wilkinson et al., 2003). The same study explored inter-rater reliability, finding that involvement in the station construction correlated well, indicating that the station writers understood what they expected from the station (2003, p. 221). Reznick and colleagues advise that a whole day is required to develop a single OSCE station (1993b). More detailed descriptions of the significant human resources requirements for writing OSCE stations was covered by Casey and colleagues (2009) as seen in Table 2-3.

Table 2-3: Representative time commitments for key OSCE personnel (Casey et al., 2009, p. 29)

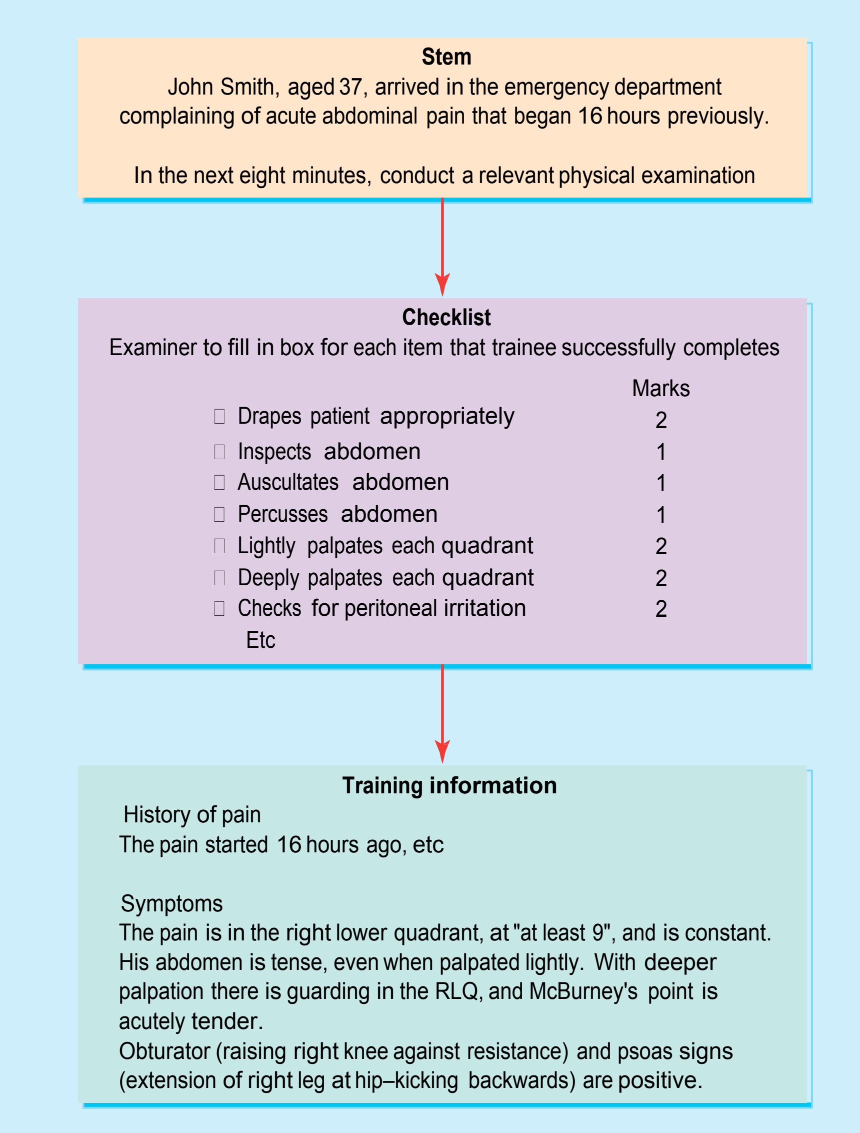
Staff role	Key training aspects	Time per 8-case examination, hours	Time used for specific tasks
Program director	Faculty educator	70	**Case review/selection**, checklist preparation, on-site attendance for examination, faculty debriefing, final scoring
		75	Preparation and teaching case-writing workshop
Faculty	Faculty educator	56	**Case writing**, pilot sessions, on-site attendance for examination
		8	**Case-writing workshop** (once)
Program coordinator	Administrative personnel	165	Coordinating examination, case preparation, on-site attendance for examination
SP coordinator	Allied health and/or performing arts background	220	SP training, **case review,** pilot sessions, on-site attendance for examination, props and makeup
		15	Teaching IPS workshop
SP	Various levels of training	60	Case preparation, including examination
		15	Attendance IPS workshop (once)

2.3.7 What is known about the station-level design principles?

The original article by Harden and colleagues (1975) included significant details regarding the structure of the examination, as well as the content of each of the stations through which the candidates rotated, each being observed performing the specified tasks. Recognition of the contribution of items, cases and simulated patient behaviour was found in more than one journal article including works by Iramaneerat and Smee (2008; 2003). Reznick, outlining the setting up of the Canadian Licensure and Certification Clinical Examination (OSCE) reports that the Medical Council of Canada 'placed a great deal of emphasis on station development' (1993a, p. s5) but supplies no further details on this critical step of the process.

Smee (2003) advises that 'stations are the backbone of an OSCE' but admits that 'station materials are incomplete and subject to last minute changes' (p. 704). Smee (2003) also describes in detail components of the written OSCE item; the stem (otherwise known as candidate or student instructions), the checklist (marking sheet) and the training information or simulated patient instruction. Whilst not providing the level of communication to candidate, examiner or simulated patient considered acceptable for station communication, an example of these components was displayed in the article and is reproduced in Figure 2.4.

Figure 2.4: Components of OSCE station (Smee, 2003, p. 704)



Stem

John Smith, aged 37, arrived in the emergency department complaining of acute abdominal pain that began 16 hours previously.

In the next eight minutes, conduct a relevant physical examination

Checklist

Examiner to fill in box for each item that trainee successfully completes

	Marks
□ Drapes patient appropriately	2
□ Inspects abdomen	1
□ Auscultates abdomen	1
□ Percusses abdomen	1
□ Lightly palpates each quadrant	2
□ Deeply palpates each quadrant	2
□ Checks for peritoneal irritation	2
Etc	

Training information

History of pain
The pain started 16 hours ago, etc

Symptoms
The pain is in the right lower quadrant, at "at least 9", and is constant. His abdomen is tense, even when palpated lightly. With deeper palpation there is guarding in the RLQ, and McBurney's point is acutely tender.
Obturator (raising right knee against resistance) and psoas signs (extension of right leg at hip–kicking backwards) are positive.

2.3.8 Language in medicine

The language used in creating stations for the OSCE writing is crucial to the communication of the task or tasks to the candidate. The consensus statement regarding assessment of performance from the Ottawa conference refers to a need for ‘consensus around use and abuse of terminology’ (Boursicot et al., 2011, p. 380). Terminology is defined by Kao as a ‘specific technical term in communication messages for highlighting the exclusive superiority of something advocated’ (2013, p. 2008). Researchers studying terminology and definitions relating to compliance and other key health related outcomes have called for ‘general and operationally useful definitions’ (Cramer et al., 2008, p. 44). Bleakley (2003) reminds us that ‘what we say to each other and how we say it matters enormously’(p. 186). Edler and Fanning argue there is an ‘imperative for standardization of nomenclature in the area of educational assessment’ and that ‘medical educators are in need of standard understanding of the terms used in educational assessment’ (2007, p. 2237). Roberts and colleagues state that the interaction between simulated patient and candidate involves both parties using scripts, with the candidate selecting from the set of medical scripts (2003). For the OSCE candidate to choose the correct script or scripts, there must be communication of the correct task through the candidate’s instructions.

2.3.9 International Medical Graduates - OSCE language and cultural bias

The globalization of the medical workforce places many International Medical Graduates in the role of OSCE candidate (Esmail & Roberts, 2013; McManus & Wakeford, 2014; Norcini et al., 2010). Broadfoot and Black (2004) in their seminal article reflecting the first ten years of the *Assessment in Education* journal, reflect on four key themes: 1) globalization, 2) purposes of assessment, 3) quality issues, and 4) assessment for learning (p. 10). Specifically, they identified key aspects affecting student performance in assessment, including student factors such as anxiety, motivation, and the use of language (Broadfoot & Black, 2004). Given the high-stakes OSCEs faced by many International Medical Graduates, with acknowledged cultural biases against doctors from non-English speaking backgrounds, language and anxiety pose major barriers to performance in this setting (Christie et al., 2011; Esmail & Roberts, 2013).

The use of a defined language or consistent terminology is advocated for in simulation to provide clarity of communication (Fairhurst et al., 2011). The dominant language or lingua franca of medical education, in an internationally mobile workforce, is English (Kane, 2014; Nestel, 2013). Patton states: ‘Language matters. It simultaneously suggests possibilities and communicates boundaries’ (1994, p. 311). Cramer and colleagues argued that an agreed language was needed for collaboration and effective use of data between different researchers (2008). OSCE station writers require an accepted language in the tasks of medical practice to effectively assess the competence of a global workforce. Without a clear dialogue about the shared understanding of OSCE task terminology, such as what constitutes a focused history, there will be a need for each institution to teach to the OSCE task as locally defined, rather than teach the skills and allow the assessment process to observe the task being performed authentically. The validity of the OSCE is undermined where a candidate is unable to understand the language of the task and thereby fails to demonstrate his or her competence in the assessment.

2.4 Psychometric Principles in Evaluation of Assessment and OSCEs

The second discourse identified in Hodges' thesis, the topic of psychometric principles, is found extensively in the literature both pre- and post-introduction of the OSCE assessment format. The purpose of the assessment and the expected interpretation of the results are important aspects affecting the validity of an OSCE. Reliability encompasses reproducibility, inter-rater and intra-rater reliability and is dependent on consistent performances by both simulated patients and examiners to standardise the experience of the candidate in the station as well as the scoring of the performance. Feasibility can be undermined when too many tasks are requested for the available timeframe. Educational impact can result from the hidden curriculum provided within the assessment, such as not providing hand-washing facilities within the examination venue, leading to an implied message regarding the perceived importance of hand hygiene.

Exploration of the key psychometric principles involved in the determination of a defendable assessment of performance was undertaken in the early stages of this literature review. This initial emphasis was necessary to better comprehend these important concepts as so much of the literature surrounding quality in the OSCE was found in the psychometric discourse. Anticipated outcomes from the literature review process included recognition of the steps and indicators for developing good assessment and improving my own assessment literacy. Concepts such as validity were extensively researched, along with reliability, utility and educational impact. The influence of generalizability theory was identified within the search domains, but does not form a significant role in this thesis due to the emphasis on improvement of the qualitative aspects of the assessment not the quantitative elements of the OSCE.

Whilst the awareness of station-level flaws and their impact on the quality of the OSCE as an assessment method was identified within the literature (Gupta et al., 2010; Iramaneerat et al., 2008; Newble, 2004), no articles other than the Heidelberg one actually provided advice on how to remedy this problem. Further expansion on the themes of the OSCE, psychometric principles, quality improvement and errors relating to OSCE item writing is presented next.

The psychometric approach to OSCE evaluation met with criticism from Schuwirth and van der Vleuten, who argued that many of the principles underlying the statistical concepts were based on flawed assumptions (Schuwirth & van der Vleuten, 2006). These assumptions relate to the transference of models of validity and reliability from the domain of psychology where traits are inherently stable, unlike the concept of competence that is contextually specific. In the OSCE format, the observation of the candidate performing the task is marked by an observing examiner who provides a score. Competence is a construct and therefore cannot be seen directly. The score provided is a combination of the true score and error; significant variation in the construct being assessed is consequently attributed to measurement error (De Champlain, 2010). A statistical concept related to classical test theory, the true score is the score expected in a perfect test, one with no errors and where the candidate could repeat the same test infinitely (De Champlain, 2010). In other words, the degree of variation in the psychometric analysis of the quality of a test is attributed to issues with the performance of the test, and not due to variation in performance on different tasks (van der Vleuten et al., 2010). An example of this variation may be the ability of an individual to perform well in history taking OSCE stations and poorly in physical examination stations (Schuwirth & van der Vleuten, 2006). Traditional psychometric approaches assume a good candidate will perform well on all tasks. Measurement error introduced through station-level flaws e.g. poor wording of the task or inadequate time provided to perform a series of tasks results in an invalid and unreliable assessment as the observed score deviates away from the true score or a true understanding of a candidate's competence.

Understanding the psychometric principles relating to the assessment of clinical competence is necessary to fully comprehend the impact on the interpretation of the results of an assessment. The key values discussed in relation to OSCE in the literature will be discussed in the next few sections of this literature review. To determine whether the quality of a station is improving following an intervention into the station writing process, it is necessary to become familiar with the terms validity, reliability, utility, educational impact and feasibility.

2.4.1 Validity

Validity is one of the key concepts used to determine the quality of the OSCE. Many assumptions are made about the ability of an assessment to predict future performance, performance in another context, performance in other tasks not witnessed, or the ability to accurately rank students in terms of their ability. Kane's discourse (2001) on the history of the concept of validity during the twentieth century relied on a definition created by Messick (1989). Messick explained that validity was the degree to which a decision, based on a score created as a result of testing, was supported by evidence, and was adequate and appropriate (Kane, 2001). To determine whether a test is valid requires knowledge of the purpose of the assessment, how the results are interpreted, and what assumptions are made based on the score resulting from the observed performance. 'Validity involves an evaluation of the overall plausibility of a proposed interpretation or use of test scores' (Kane, 2001, p. 328). According to Kane's definition of validity, 'it is the interpretation (including inferences and decisions) that is validated, not the test or the test score' (2001, p. 328).

2.4.2 Reliability

Reliability is another psychometric principle used to determine the quality of an OSCE. A sound assessment should be both valid and reliable (Kaslow et al., 2007; van der Vleuten, 2000). Interpreting a reliable instrument as having little variation in results on repeated testing is a concept often illustrated using the example of sphygmomanometry or blood pressure readings (Cook & Beckman, 2006). Norman (2014) explains that reliability is the 'ability of a measurement instrument to consistently discriminate' between having abundance or scarcity of a desired characteristic (p. 946). Norman (2014) admits to the awkwardness of this interpretation and expands on the concept of reliability being the degree of variance that can be attributed to a real difference in OSCE performance, not other factors, such as the effect of different raters (examiners) or undertaking the test at different times. These factors are known in mathematics as error. Whilst reliability is a necessary element for validity, validity requires more than reliability to provide sufficient evidence to justify the interpretation of the score (Downing, 2004). Significant publication numbers revolved around the aim to improve reliability, however the failure to recognise that reliability alone does not equal validity is a potential flaw in these papers. In an OSCE setting, the concept of reliability using the Taylorism discourse involves standardisation and training of examiners and simulated patients and the station environment to ensure a uniformity of experience for candidates. As discussed in the introduction to psychometric principles, an OSCE with high reliability should determine that the variance in a candidate's score is due to the ability of the candidate to perform the task and not due to other aspects of the examination.

2.4.4 Utility

'Utility', according to Broome, 'means usefulness' (1991, p. 1). He expands on the history of the term from both an economic and a philosophical perspective quoting Bentham and Mill before arriving at his preferred definition of 'that which represents a person's preferences' (Broome, 1991, p. 11). In the economic sense of the word, a choice is made between two functions; in medical education, utility is interpreted as the beneficial functions attributed to an action, activity, or intervention (Bamber et al., 2014). Both validity and reliability were included in van der Vleuten's utility formula that combined five variables: reliability, validity, educational impact, acceptability and cost to create a model for assessment decisions (1996). The original utility formula has been represented by the following equation where R= Reliability, V= Validity, E= Educational impact, A= Acceptability and C= Cost (Chandratilake et al., 2010, p. 6):

$$Utility = R \times V \times E \times A \times C$$

Chandratilake and colleagues (2010 p. 7) proposed a new utility formula incorporating feasibility as part of the equation and replacing cost with cost-effectiveness where R= Reliability, V= Validity, EI= Educational impact, P = Practicability, A= Acceptability and CE= Cost- effectiveness:

$$Utility = R \times V \times EI \times P \times A \times CE$$

In this version of the utility formula, practicability was synonymous with feasibility, and the revelation that if any one element of the equation was absent or zero then the utility of an assessment would also be zero (Chandratilake et al., 2010). In this endeavour to advance OSCE quality through station-level writing improvements, the utility of the assessment using van der Vleuten's 1996 model, where attention is focused on aspects which may affect reliability, validity, educational impact, acceptability and cost, is relevant. However, the inclusion of feasibility, in the context of station-level errors, was also explored.

2.4.5 Educational impact

A significant element derived from OSCE station content is the influence on future student behaviour (Hodges, 2003a). Hodges argues, 'the OSCE has led to much more attention being given to the performance of certain professional behaviours, including patient-centered interviewing, cross-cultural competence and interprofessional communication' (2003a, p. 253). Educational impact derives from the accepted and observed phenomenon that assessment drives learning behaviour in students (Newble & Jaeger, 1983; van der Vleuten, 1996). Positive and negative impacts are possible; the effect of checklists on student performance in OSCE has been well documented – the student approach to learning is not to practice the skill but to memorise the checklists (Cunnington et al., 1997).

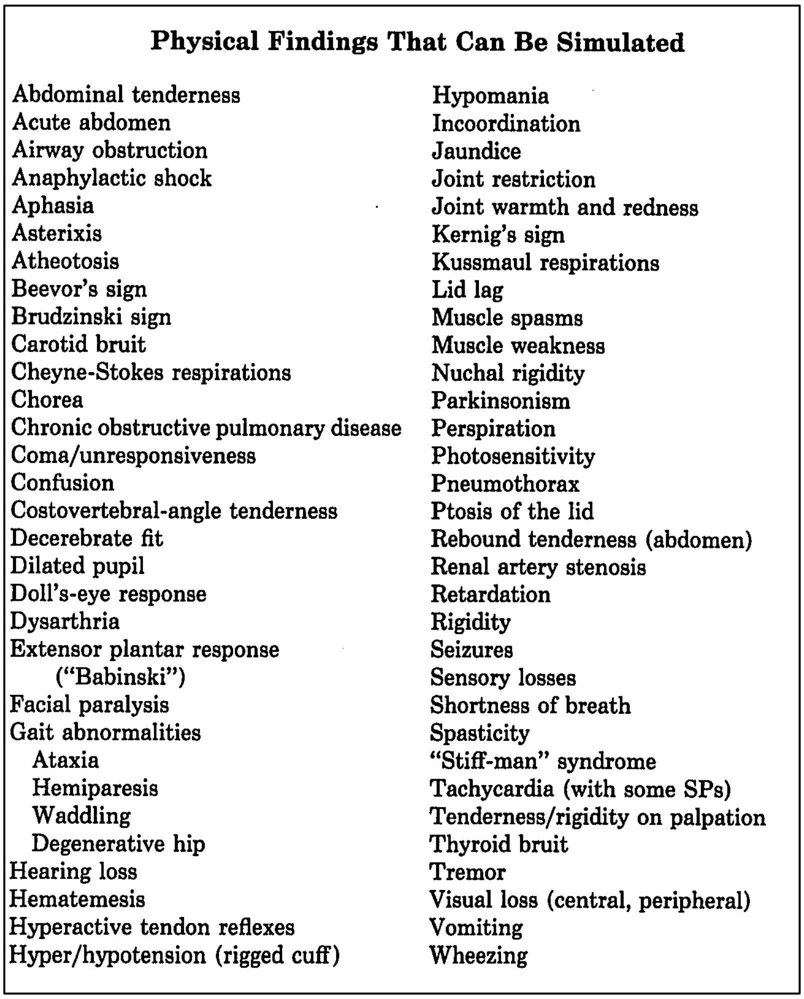
2.4.6 Feasibility

None of the medical education papers I reviewed with this word in the title or keywords section defined this term within the publication, indicating an assumed shared understanding of this term. Feasibility is defined by the Oxford dictionary as the 'state of being easily or conveniently done' (Oxford Dictionaries). Feasibility was assessed in a number of papers with the costs of running an OSCE being perceived as a particular barrier to the adoption of this format of assessment (Barman, 2005; Eberhard et al., 2011; Hingle et al., 2011; Poenaru et al., 1997). Reznick and colleagues published a guide to estimating the true cost of OSCEs based on the Canadian national licensing examination in addition to provincial and institutional OSCE experiences (Reznick et al., 1993b). OSCE scenario and station-writing by teams in whole-day workshops was considered more productive than individual instruction and independent case-writing (Reznick et al., 1993b). From the administration of the assessment, feasibility or practicability was considered in relation to the use of scanners or electronic scoring to improve efficiency and decrease mistakes from manual scoring (Barman, 2005). Yet, despite criticisms of the resource requirements, the ability to assess large numbers of students in a standardised, fair and acceptable format still led to the OSCE as the preferred approach (Eberhard et al., 2011; Pell et al., 2013).

Other aspects of feasibility are relevant to OSCE station writing. The time allowed for the station creates limitations on the type of task and the number of tasks a candidate can reasonably be able to undertake without compromising safety, patient-centred medicine, or creating a situation where the candidate is forced to take short-cuts in one or more tasks to complete the station. According to Khan and colleagues, 'an appropriate and realistic time allocation for tasks at individual stations will improve the test validity' (2013a, p. e1449).

Feasibility was also considered when the choice of marking sheet style, global rating or checklists was considered from the perspective of the examiner (Ringsted et al., 2003). Decades of research have expanded the concepts of what may be feasible to assessing using simulated patients and the OSCE format (Barrows, 1993). An early collation of simulated patient capabilities, Barrow's list of what physical traits a simulated patient could portray in 1993 is reproduced in Figure 2-5.

Figure 2-5: Barrows' list of physical findings that can be simulated. (1993, p. 445)



Physical Findings That Can Be Simulated

Abdominal tenderness
Acute abdomen
Airway obstruction
Anaphylactic shock
Aphasia
Asterixis
Atheotosis
Beevor's sign
Brudzinski sign
Carotid bruit
Cheyne-Stokes respirations
Chorea
Chronic obstructive pulmonary disease
Coma/unresponsiveness
Confusion
Costovertebral-angle tenderness
Decerebrate fit
Dilated pupil
Doll's-eye response
Dysarthria
Extensor plantar response
("Babinski")
Facial paralysis
Gait abnormalities
Ataxia
Hemiparesis
Waddling
Degenerative hip
Hearing loss
Hematemesis
Hyperactive tendon reflexes
Hyper/hypotension (rigged cuff)

Hypomania
Incoordination
Jaundice
Joint restriction
Joint warmth and redness
Kernig's sign
Kussmaul respirations
Lid lag
Muscle spasms
Muscle weakness
Nuchal rigidity
Parkinsonism
Perspiration
Photosensitivity
Pneumothorax
Ptosis of the lid
Rebound tenderness (abdomen)
Renal artery stenosis
Retardation
Rigidity
Seizures
Sensory losses
Shortness of breath
Spasticity
"Stiff-man" syndrome
Tachycardia (with some SPs)
Tenderness/rigidity on palpation
Thyroid bruit
Tremor
Visual loss (central, peripheral)
Vomiting
Wheezing

Unfortunately, whilst Khan et al. (2013a) focus on the training and selection of simulated patients to ensure station standardisation, the role of the clarity required for the simulated patient instructions is not discussed.

Decisions concerning station content are important considerations in the feasibility of the proposed assessment. Gormley and colleagues describe the factors to consider in an 'ideal encounter with a [simulated] patient in an objective structured clinical examination' (2012). A task should meet the following criteria, that: a) a simulated patient would be willing to participate, b) it does not pose a risk for the physical or mental well-being of a simulated patient, c) is not too complex for the purposes of training the simulated patient, d) it does not require highly complex equipment, e) the set up or reset time in the station is achievable in the time provided, and f) equipment used has a low risk of failure or takes minimal time for backup equipment to be set up (Gormley et al., 2012, p. 383). These factors are important considerations in OSCE station writing and quality assurance processes for improving simulation based assessment.

2.5 Quality Processes in Assessment

Quality assurance is considered essential in creating ‘fair, rigorous decision making about candidates’ in high-stakes testing using OSCE, including maintaining an institution’s reputation (Fuller et al., 2013, p. 515; Pell et al., 2013, p. 515). Quality improvement has also been advocated as part of the curriculum for medical students to improve patient safety and care (Wong et al., 2012). Predominantly, assessment of quality has involved the use of classical test theory to measure the reliability of the assessment using statistical items such as Cronbach’s Alpha which measures internal consistency (Eberhard et al., 2011; Tavakol & Dennick, 2012).

The need to ensure that the OSCE item, identified within this thesis as the OSCE station is of a high quality is still important, even if a cautious approach to the use of item analysis is advised (van der Vleuten & Swanson, 1990; Yudkowsky et al., 2014). Schuwirth and van der Vleuten were highly critical of the approach to quality improvement that involved the removal of poorly performing items following item-based analysis, despite the possibility that they might under scrutiny be ‘found to be relevant, correctly phrased, part of the objectives of the course, taught correctly and had content beyond doubt’ (2006 p298).

2.5.1 The consequences of OSCE errors

Errors exist in the OSCE setting, some of which lie within the station or item content (Vallevand, 2008). Consequences resulting from flawed clinical assessments can be considered as two types of statistical error. Failing the good or competent candidate due to assessment structure or process defects is regarded as a Type I statistical error (Crichton, 1998). Type II statistical errors are those where a candidate who is not competent or meeting the required standard is passed due to a faulty assessment (Crichton, 1998). Both error types may be seen where assessments deviate from known best practice. Errors made in either the creative content or the communication aspects of the assessment can undermine the quality of the assessment. Exploring the concept of errors and error prevention in OSCE station writing to improve the quality of assessing clinical skills is fundamental to this research project.

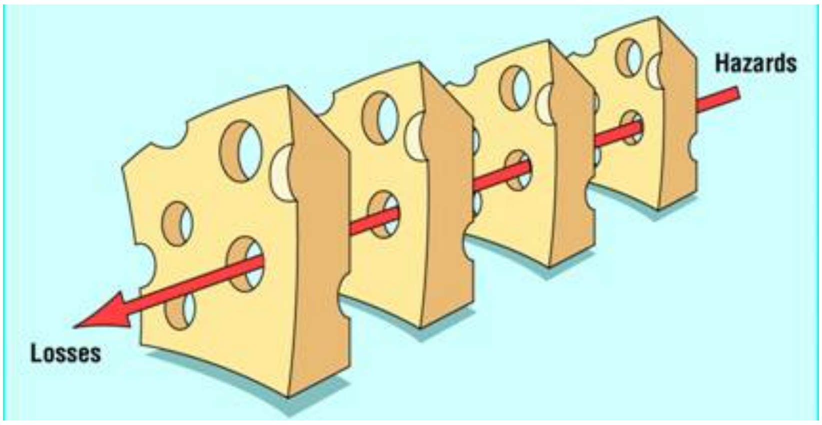
Within the OSCE quality assessment framework, item analysis is used to determine which stations should be eliminated from the final examination results due to poor item statistics. Cronbach’s alpha, a measure of internal consistency, is used in OSCE where the station results are combined to give an overall examination result (Brailovsky & Grand'Maison, 2000). For the purposes of combining results, an assumption is made that the items are all measuring the same thing, i.e. they have internal consistency (Bland & Altman, 1977). In OSCE, we aim for results with an alpha approaching 0.8 to 0.9 suggesting we are testing a similar but not identical construct (Boursicot et al., 2006). If the removal of a station leads to a higher Cronbach's alpha then it is considered to be a flawed station; however, Schuwirth and van der Vleuten caution against a reduction in the sample size through removal of an otherwise acceptable station (Schuwirth & van der Vleuten, 2006). The acceptance of low Cronbach's alphas is of concern where decisions are made on the basis of the test results (Schuwirth & van der Vleuten, 2006). The fact that OSCEs contain stations of different types encourages this attitude, as does the fear of reduced validity from the smaller sample.

Another possibility exists, that there is a normalisation of deviance within the medical assessment fraternity (Kan Ma et al., 2013). The concept of a lowering of the tolerance to substandard performance is a component of the human factors aspects explored in the interests of patient safety (Banja, 2010). We accept a lower standard in the belief that when the standard was lowered previously, no adverse event occurred, therefore, lowering the standard further will have the same outcome. This is particularly easy to tolerate in assessment where the passing of the incompetent student will be harder to detect than the death of a patient from a medication error. There are many other parallels with the patient safety framework that could be explored with relevance to the introduction and failure to detect errors within the OSCE. A key concept discussed with respect to aviation, and more recently adopted by medicine, includes a heighted awareness of the likelihood that errors will occur, and therefore placing systems in situ to detect these errors before harm eventuates (Banja, 2010).

2.5.2 Systematic approach to errors

Undertaking a systematic approach to detection and correction of errors is a pillar of the patient safety framework (Noble & Donaldson, 2011) and has also been applied successfully in the setting of quality improvement in the OSCE (Schultz et al., 2008). Error is defined in the landmark *To Err is Human* report on safety in health care as 'the failure of a planned action to be completed as intended (e.g. error of execution) or the use of a wrong plan to achieve an aim (e.g. error of planning)' (Kohn et al., 2000, p. 54). The further classification of errors as active (immediate impact) or latent (more indirect), identifies poor design as one source of latent errors (Kohn et al., 2000). Undertaking a systematic approach to error detection and prevention has been advocated to reduce adverse events in the hospital setting and in primary care (Singh et al., 2014; Thomas et al., 2011). James Reason's Swiss cheese model for understanding the need for a systematic approach to error prevention is well recognised and cited (Reason, 2000). This model is shown in Figure 2-6.

Figure 2.6: The Swiss cheese model of how defences, barriers, and safeguards may be penetrated by an accident trajectory (Reason, 2000, p. 769)



The imagery of this diagram evokes an understanding that for an error to persist through to causing injury requires the failure of multiple opportunities to prevent this occurrence. This assumes that such safeguards are in place. In assessment, Van der Vleuten and colleagues write that 'Quality appraisal of tests during the developmental stage is imperative' (2010, p. 5). In addition, they advise that 'peer review is an essential ingredient' of quality improvement systems to improve assessment materials e.g. OSCE stations (van der Vleuten et al., 2010, p. 5).

2.5.3 Teamwork and Peer Review

The use of faculty teams to review newly created or proposed recycled OSCE stations prior to the examination with the provision of a checklist to aid this process is supported by a single study as discussed in the introduction to this chapter (Schultz et al., 2008). Along with faculty development for examiners and the use of psychometric analysis to provide feedback directly to station writers, the introduction of the OSCE station review checklist in Heidelberg resulted in an improvement in the reliability of the assessment as measured by the internal consistency statistic, Cronbach's alpha (Cohen et al., 1990; Schultz et al., 2008, p. 672). The checklist was used at OSCE review meetings attended by faculty from all the medical schools in Germany. Teamwork is already identified as an element of the new curriculum along with human factors and situational awareness training as a response to the well-documented tragedy of human morbidity and mortality through preventable errors in the hospital setting (Donaldson, 2009; Gawron et al., 2006; Kohn et al., 2000). In the setting of OSCE item writing, actions advocated for patient safety may also be constructive; including teamwork, accepting the inevitability of errors and adopting a system-based approach to improve OSCE quality. This approach was modelled by the assessment team led by the University of Heidelberg faculty (Schultz et al., 2008). A copy of the Heidelberg checklist, including a translated version is provided in Appendix A.

2.5.4 Assessment literacy

Assessment literacy is an important component of a competent educator's skill set. Popham's (2009, 2011) work from the field of education provided the best resource for understanding the concept of assessment literacy (Popham, 2009). Defined as 'an individual's understanding of the fundamental assessment concepts and procedures deemed likely to influence educational decisions,' assessment literacy has relevance in health professional education (Popham, 2011, p. 265). Assessment literacy of students in higher education has been considered significant, particularly the concept of first year students understanding the purpose of assessment (Smith et al., 2011). The assessment literacy of clinicians has not been well documented, but is assumed to be low due to a lack of emphasis on education qualifications in the medical education domain (Eitel et al., 2000). Cook and Beckman (2006) reported on the poor comprehension of the true meaning of validity and reliability by clinical teaching physicians. This lack of understanding of basic assessment terminology supports the concept of assessment literacy as a desirable trait in medical educators involved in creating assessment items.

2.6 Writing for OSCE

Writing stations for an OSCE is a time consuming task, fraught with potential misjudgements regarding candidate, simulated patient (SP) or logistical capabilities (Hettinga et al., 2010; Vargas et al., 2007). Station or case-related issues include 'those associated with SP portrayal, unanticipated student reactions to the scripted SP responses, and case irregularities (e.g. patient history and/or physical findings are not consistent with the intended diagnoses)' (Vargas et al., 2007, p. 194). Van der Vleuten and Swanson, commenting on the potential impact of a mismatch between the station writer's intended tasks and the perception of the examinee, dubbed this 'the guess-what-I-want-you-to-do problem' (1990, p. 72).

For most institutions the luxury of a case-writing team that includes 'health professionals, SP trainers, and educational experts/psychometricians' is not achievable (King et al., 1994,

p. 8). Vargus and colleagues report that ‘case development is an iterative process; until the scenario is acted out, and some pilot administrations undertaken, it is difficult to discern all the potential problems’ (2007, p. 194).

Faculty development in station or case writing has been shown to improve the quality of the OSCE (Schultz et al., 2008). Station-writing training has also been demonstrated to improve the quality of both written and OSCE test items as measured using repeated analysis of variance measurements pre- and post- faculty development (Naeem et al., 2012). Vargus and colleagues were positive about the effect of faculty development at their institution: ‘Faculty development efforts at NUC in the field of educational measurement and assessment have certainly led to a more equitable and defensible examination’ (2007, p. 196). Holmboe and colleagues (2011) emphasised deficiencies in the ability of faculty to effectively assess performance of trainees. However, their five necessary steps to improve faculty assessment skills do not include the ability to write valid and reliable assessment items such as OSCE stations (Holmboe et al., 2011).

Whilst some faculty development clearly exists in some institutions, in others the process of station writing falls to clinicians with little or no training in OSCE station development (van der Vleuten et al., 2010). According to van der Vleuten and colleagues, ‘it is not uncommon for test materials in medical schools to go unreviewed both before and after test administration’ (2010, p. 5). Given that Boulet and colleagues identified that the ‘interaction between examinees and cases is a major source of measurement error’ (1998, p. 91) a systematic approach to identifying errors at the station writing level is warranted.

2.6.1 Flaws in OSCE station-level writing

Reflecting on the challenges in current assessment practices, Epstein states that educators ‘should be mindful of the impact of assessment on learning, the potential unintended effects of assessment, the limitations of each method (including cost)’ (Epstein, 2007, p. 394). Varkey and colleagues admit to the need for changes in OSCE stations following a pilot process of newly created stations for assessing quality improvement (2008). Interestingly, the faculty in their study modelled an effective quality improvement process beginning with the use of an OSCE writing committee, then piloting their new stations with a different group of clinicians who had not been involved in the writing process (Varkey et al., 2008). Iramaneerat and colleagues (2008) identified that the largest source of variance in their study was due to simulated patient/case effects, but were unable to demonstrate the effect of the case alone. A study exploring the effect of removal of problem stations in

Thailand identified 33/217 problem stations in examinations for eight cohorts of medical students (Auewarakul et al., 2005). Specific errors in the writing were not discussed in detail in this study although 'remediation of the station' was recommended along with changes to the teaching of the pathology skills relating to the bulk of problem stations (Auewarakul et al., 2005, p. 112).

2.7 Chapter conclusion

Although hundreds of publications relate specifically to the use of the assessment format OSCE, only one provided a solution designed to improve the quality of the OSCE through discovery and remediation of station-level errors. Published in a German language journal it is unlikely that this reference has had the recognition that might have occurred if the full study was published in English, and not just the abstract contents. Other publications acknowledged the importance of station-level design flaws through their effect on reliability or validity. The importance of assessment driving learning or educational impact of the OSCE is well recognised in the literature, and some specific details regarding threats to feasibility were defined. Whilst the Heidelberg checklist provided a similar solution to that proposed in this project, it did not address the issues of educational impact, or feasibility. This project will assist OSCE station writers to reflect on station content that has the potential to undermine not just the validity or reliability, but also the feasibility or educational impact of these errors. Ultimately by doing so, the station writers can identify and remediate issues to improve the quality of the clinical assessment.

This chapter explored the literature pertaining to the assessment of competency with particular emphasis on the format known as the Objective Structured Clinical Examination or OSCE. The literature relating to the history of the OSCE, the use of psychometric principles to analyse the quality of an OSCE, the quality improvement processes involved in the OSCE and the writing and development phase of the OSCE were included. The chapter explored the approach to the literature, key outcomes and the relevance to the project of specific examples from the vast literature available on this topic. The following chapter will explore the research methodology, and specifically the method used in this thesis project along with the approach taken with respect to ethics.

Chapter 3 – Methods

3.1 Introduction

This chapter identifies the methods used in this project and provides the background for the choice of research methodology, the steps involved in gathering data, and the ethical considerations raised by this project. The use of design-based research methodology to aid in solving a problem relating to assessment of clinical competence will be described.

The aim of this research is on improving the quality of the assessment of clinical competence using the Objective Structured Clinical Examination format, by aiding station developers and reviewers to identify station-level errors. This project was born out of the observation of failure to recognise flaws in the design of OSCE stations. These observations occurred over time and in multiple educational contexts. Understanding what constituted the best practices in OSCE station writing required careful exploration of the available literature, and included personal discussions with key medical assessment researchers across many countries. Translation of this research knowledge into action was an underlying concern that required research methods able to support this aim.

Ringsted and colleagues advise that a conceptual, theoretical framework is required to move beyond an idea or problem to a research project (Ringsted et al., 2011). The research question is a key consideration in the choice of method or approach. A series of small reflective steps beginning with the actual problem is undertaken, to determine the research question. In this project the following research question was posed:

> *What aspects of the OSCE item writing process are prone to errors that undermine the quality of this assessment format and how can these be overcome?*

However, given the iterative nature of design-based research, this question evolved along the journey of this project. Rather than a specific question, a series of questions emerged. These questions included the following:

1. *What elements of known best practice in OSCE station writing should be included in a tool to aid OSCE writers and reviewers to improve the performance of the OSCE?*

2. *What steps are necessary to create a useful tool to assist OSCE writers and reviewers to identify potential errors in the writing phase of the assessment process?*

3. *Does the tool make the invisible visible with respect to flaws within an OSCE station, enabling writers and reviewers to identify errors during the pre-exam quality improvement processes?*

4. *What is the utility of a tool to enhance the quality of OSCE stations in assessment?*

Observation of a candidate's performance to assess competence within a constructed simulated clinical examination is a complex undertaking and errors arise during station development. Given the complexity of this problem, this project has an aim rather than a specific research question. To achieve the aim of this research, the intent became to create a tool to aid OSCE station writers and reviewers to identify flaws affecting the ability of the candidate to perform the task. The tool will also provide a mechanism to explore those aspects of a station that enable multiple circuits to provide the same experience for all candidates and to ensure that those aspects of the examination that might drive future clinical behaviours of candidates and examiners are evidence-based.

Development of a tool to address the observed problem of poorly-performing OSCEs, due to errors in the written content, lies within the descriptive studies domain of research requiring more than a description of the innovation to qualify as research (Bannan-Ritland, 2003; Dolmans & van der Vleuten, 2010). Ringsted and colleagues concede that if a descriptive study 'addresses a research question that relates to a conceptual, theoretical framework, it stands a better chance of being accepted as research' (2011, p. 698). The conceptual theoretical framework involves three elements: 1) identifying which theories of learning and education can aid understanding of the problem or idea, 2) critical exploration of current research knowledge, and theories to position the research and 3) the novelty of the contribution from the researcher on this topic (Ringsted et al., 2011).

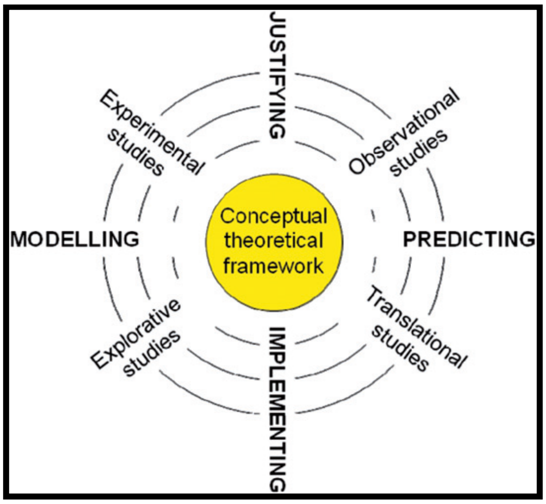
Schuwirth and van der Vleuten (2011a) offered an overview of the theoretical framework underpinning assessment in medical education, providing insight into contextual relevance of my project within currently accepted views. Their exploration of psychometric theories relating to reliability and validity reveal impediments in the interpretation of these terms, whilst the discussion around the different classes of test theories to determine reliability highlights the complex nature of this field (Schuwirth & van der Vleuten, 2011a). Prideaux (2002) also highlighted the need for the application of a conceptual theoretical framework to improve the robustness of medical education research. Understanding the scholarly relevance of this research requires consideration of three key elements: 1) the mechanisms underlying the development of errors, 2) the literature review exploring the topic of OSCE and errors (including the psychometric parameters relating to the quality of OSCE), and finally, 3) the creative process employed to address the research question. These elements are reproduced in Table 3-1 Application of the conceptual theoretical framework.

Table 3-1: Application of the theoretical conceptual framework

Conceptual Theoretical Framework Elements (Ringsted et al., 2011)	**Application of Theoretical Conceptual Framework to research project**
Identification of relevant theories of learning and education to aid understanding of the problem or idea	Consideration of the mechanisms underlying the development of errors e.g. assessment literacy, faculty development in station-writing
Critical exploration of current research knowledge and theories to position the research	Literature review exploring the topic of OSCE and errors (including the psychometric parameters relating to the quality of OSCE)
The novelty of the contribution from the researcher on this topic	The creative process employed to address the research question e.g. development of a tool to aid OSCE writers and reviewers to identify station-level errors.

According to Ringsted and colleagues (2011) descriptive studies fall under the exploratory studies zone on their Research Compass model. Use of this model illustrates where this research project lies with respect to research methodology approaches and positions it for interpretation in the context of existing research on this topic. By linking the published research with OSCE station-level errors, this research falls between modelling and implementing, on the model depicted in Figure 3-1.

Figure 3-1: The Research Compass. An overview model of approaches to research in medical education (Ringsted et al., 2011, p. 698).



An appraisal of the available literature was undertaken to identify the most appropriate methodology for this project. A practical approach was desirable to address concerns about assessment quality. Consideration of emerging research methods on the recommendation of my supervisors stimulated interest in exploring alternatives to traditional methods.

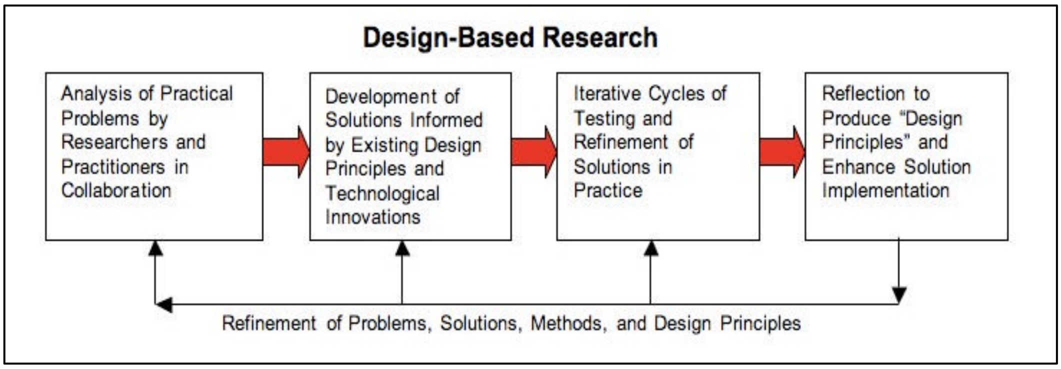
These traditional approaches to research are usually grouped into qualitative and quantitative methods (Hopper, 2008). Quantitative research, according to Tavakol and Sandars, starts with the creation of a hypothesis based on known scientific theories, then tests this hypothesis by gathering data and measuring the effect size using validated tools (2014). Quantitative research is used to compare and contrast two populations or interventions and can answer questions relating to who undertook a particular activity or what happened in the realm of research activity (Given, 2008). Qualitative research explores the how and why questions using descriptive language and explorative or interpretive approaches (Given, 2008). Typically, qualitative research methods involve getting closer to the research subject in search of the answer to how or why things work (Hopper, 2008). Morse defines qualitative health research as an approach to investigating health and illness from the perspective of the subject person or population, rather than from the researcher's opinion (2012). Neither quantitative nor qualitative methodology exclusively, as recognised within the traditional scientific model of research design, provided the necessary practical application to address the identified educational problem. Limited to defining, describing or providing evidence relating to the existence of the errors observed within OSCE stations, traditional qualitative and quantitative methodology lacked the potential to meet the aims of this thesis.

3.2 Research method

One method described in the literature which purports to address issues of transformation in research is known as 'education design research' or 'design-based research' (Brown, 1992). For the purpose of consistency throughout this thesis, the term design-based research will be used. In design-based research the process involves identifying a possible solution for a local problem through review of the literature or application of a theory, and focusing on both the design process and the outcome (Barab & Squire, 2004). The research designed to mitigate a problem starts off in a local context, and if positive results emerge should be able to translate to improved outcomes in other learning settings (Dede, 2005).

In the structure of design-based research, four stages are typically described as being integral to the research process (Anderson & Shattuck, 2012). These stages consist of the following: 1) identification of a problem in a local context, 2) designing a potential solution and testing the outcomes of the intervention, 3) incorporating methods from all potential research methodologies with no allegiance to any particular approach, and 4) the use of an iterative process to improve the intervention following testing in the education setting (Anderson & Shattuck, 2012, pp. 16-17). Design-based research can be summarised as defining a problem, identifying solutions from within the available research knowledge and applying them in field-testing in different environments to refine and define the utility of the local solution, aiming for a more global applicability. A visual interpretation of this process, illustrated by Reeves and reproduced by Herrington (2007) in a discussion paper regarding the use of design-based research in theses is shown in Figure 3-2.

Figure 3-2: Design-based research approach (Herrington et al., 2007)


Design-Based Research
Analysis of Practical Problems by Researchers and Practitioners in Collaboration
Development of Solutions Informed by Existing Design Principles and Technological Innovations
Iterative Cycles of Testing and Refinement of Solutions in Practice
Reflection to Produce "Design Principles" and Enhance Solution Implementation
Refinement of Problems, Solutions, Methods, and Design Principles

Design-based research has evolved from concerns related to the interface between educational research, and its implementation in practice. As noted in chapter 2, poor quality health-professional clinical assessments still persist, despite the expertise and published eloquence of many academics in this field. Design-based research presented a suitable method to address the research objective due to its problem-resolution focus (McKenney & Reeves, 2013). In keeping with the message behind the importance of the research compass to enable understanding of the underlying theories and principles relating to a research project (Ringsted et al., 2011), McKenney and Reeves assert that the theoretical framework associated with design-based research 'can be descriptive, explanatory or predictive in nature' (2013, p. 98).

Reeves and colleagues (2011) describe the methodology associated with design-based research in the educational setting, as beginning with identifying a problem relating to teaching and learning. As described earlier in this chapter, design-based research includes development of a solution to the problem based on existing research knowledge and principles (Reeves et al., 2011). Context is an important feature of this chosen method. A response to an intervention may work well in one setting, yet fail to have an effect in another. In design-based research, refinements are made, and the tool re-evaluated, until a satisfactory solution is obtained which may translate across different situations.

Design-based research has been used to solve problems in the medical educational setting. In 2012, Dolmans and Tigelaar advocated for the use of design-based research in medical education 'because these studies both advance the testing and refinement of theories and advance educational practice' (p. 1). Highlighting the potential benefit of this research approach in the areas of work-based learning and assessment, an example provided in the AMEE Guide number 60 included research into the successful introduction of a portfolio for encouraging teacher professional development at Maastricht University (Dolmans & Tigelaar, 2012; Tigelaar et al., 2006). Dornan and colleagues used design-based research to explore self-directed learning, including behaviour and opinion, in medical students (2005). An effectiveness study exploring the introduction of online learning modules in surgical training met the criteria for design-based research, and used the word 'design' 26 times within the publication; however, this paper was not formally identified by the authors as design-based research (Ellaway et al., 2014). Similarly, Tsai and Harasym's (2010) development of a medical ethical reasoning model did not identify the method used as design-based research, yet meets the criteria for this in the conduct of their study that described the reasoning, design process and evaluation of this intervention (2010). Another example of design-based research, Haidet and colleagues (2005) created and validated a measure for exploring the hidden curriculum in ten medical schools in the USA, using an iterative process to explore patient-centered care in undergraduate medical education. Schuwirth and van de Vleuten (2011) advocate defining and adhering to a specific methodology, including a clearly defined theoretical framework, as a crucial step, when conducting medical education research.

There have been some criticisms of design-based research. These criticisms include concerns about the validity of the method as a research process. Limitations and criticisms of design-based research methods were identified through an extensive review of the available literature on this topic including concerns about the immaturity of the methodology (Wang & Hannafin, 2005). Another criticism from Wang and Hannafin (2005), labelled 'Paradigm Shift' relates to the role of the researcher who may be embedded within the project and thereby effect the result of the research study (2005). This criticism was also stated by Anderson and Shattuck (2012) in their review on a decade of design-based education research (2012). Apart from the loss of objectivity, the researchers' own beliefs that the project will succeed may result in increased enthusiasm and willingness to work harder to achieve this outcome (Wang & Hannafin, 2005). The translatability to an alternative setting where this level of influence is no longer present may undermine the results of further studies using the same design (Akkerman et al., 2013). Likewise, the traditional application of theory into practice to resolve an issue may be unsuccessful where the theory has no relevance to the actual problem on the ground. The inability to control the environment sufficiently within the educational setting also creates challenges for the reproducibility of results obtained using design-based research.

This project explores existing research based on biases and heuristics, social cognition theory and naturalistic decision-making theories as discussed by Berendonk and colleagues (2013) in the setting of performance-based assessment. It combines both the psychometric model of competency assessment through the attempt to improve the quantitative reliability of the OSCE assessment tool, with the constructivist approach whereby the impact of the assessment is considered to have a possible negative impact on a student's learning. The acceptable use of subjective decision-making in an objective designed assessment framework (van der Vleuten et al., 1991) is still likely to benefit from exploratory studies aligned with research in biases and heuristics (Tversky & Kahneman, 1974), whereby a reduction in errors affecting the observed performance improves the validity of the results. The naturalistic decision-making research (Klein, 2008) explains the current approach of examiners faced with highly flawed OSCE stations, that they are likely to consider prior experiences and expertise to make the best possible decisions based on the available information in a compromised examination.

Qualitative studies form the bulk of the literature relating to OSCEs in medical education research. These papers explored errors and the subjective nature of judgements based on observed performance. Valentino and colleagues (1998) provided an excellent example of a qualitative research approach exploring inter-rater reliability using a questionnaire to survey faculty members' opinions regarding which items should be included in an OSCE station checklist. In contrast, literature relating to psychometrics frequently depicted the use of quantitative research methods. Sibbald & Regehr (2003), researching the feasibility of having first year pharmacy students act as simulated patients for the final year pharmacy OSCE, used quantitative research methods to analyse psychometric values. Both qualitative and quantitative research methods have been extensively applied to the OSCE creating a richness of knowledge surrounding the what, how and why of this assessment format.

Hodges (2003b), in his treatise on the OSCE, questioned the validity of the approaches used to measure validity. His discussion also included the inherent aspects of the OSCE that contribute to the validity of this approach to assessment. Reference to the OSCE as 'a very powerful tool that defines and reinforces particular behaviour' (Hodges, 2003b, p. 251) and the need to pay attention to the scripted simulated patient roles (Hodges, 2003a) demonstrated alignment with this project's aim. Reviews of available research on OSCEs are plentiful, and there are both systematic and haphazard approaches incorporated in these collations of fact and opinion. It is important, however, not to discount the importance of these contributions to the wealth of knowledge regarding what works and what does not work using the OSCE for assessment.

The existence of this dichotomy led to an appreciation that neither a quantitative research approach, nor a qualitative approach alone would be appropriate. Improving the implementation of established research outcomes in the field of clinical skills assessment, a fundamental aim of this research, required a different method from the traditional options. A practical technique to address the problem, involving the introduction of a potential solution, required a research method applicable to the medical education setting. The design-based research approach, described earlier in this chapter by Reeves and colleagues (2011), seemed suited to the aim of this study. Design-based methodology with the use of descriptive, iterative approaches to resolving a practical problem in education was therefore identified as a suitable method for this thesis project.

3.3 Ethical considerations

An application for ethical approval of research conducted for the purpose of this thesis project was submitted to the Flinders University and Southern Adelaide Health Service Social and Behavioural Research Ethics Committee (SBREC) in June 2011. Responding in the negative to questions relating to the use of human subjects, indigenous people, Commonwealth data or hospital health records, or research involving animals resulted in a response from the committee advising that ethics approval was not required for the planned research. The experience gained through the application process provided valuable research-skills training and was a necessary step in the Master's project journey.

Whilst a large collection of OSCE documents was sourced during this project from a number of institutions, these valuable assessment resources have remained de-identified for the purpose of this thesis. Collated into a large database, relevant details such as the task(s) being assessed and level of education of the target candidate were recorded. Institutional details were irrelevant for the requirements of testing the tool against a broad range and styles of OSCE writing; however, for the purpose of validation of this work, a coded system was applied to enable retrieval of specific OSCEs from within the database if required. Anonymity was crucial where stations may still be in active use as potential assessments for current students. Furthermore, access to the database was restricted to the researcher, and as such, all records remained locked in filing cabinets and on password-protected computers. Maintaining the integrity of examination materials was a key consideration of the ethical conduct of this project.

3.4 Establishing rigour / trustworthiness of study

Validation of the tool is recognised as an important component of the development of a tool; however, this thesis does not fulfil all the requirements of a fully validated tool (Hawkins et al., 2010; Ilic et al., 2014; Schou et al., 2012). According to Beckman and colleagues, the American Psychological Association and the American Educational Research Association publish a set of standards relating to evidence accepted for evaluating content validity in education assessment tools or instruments (2004). The five categories identified in the 1999 edition, for the purpose of construct validity, include the need for evidence relating to ‘1) content, 2) responses, 3) internal structures, 4) relationship to other variables, and 5) consequences’ (Beckman et al., 2004, p. 973). The validation process does not require evidence in all categories; however, the use of more than one category is recommended. The 2014 standards published by the American Educational Research Association and the American Psychological Association advise that a test is a ‘device or procedure in which a sample of an examinee’s behaviour in a specified domain is obtained and subsequently evaluated and scored using a standardised process’ or an instrument ‘on which responses are evaluated for the correctness or quality’ (2014, p. 2). The process of publishing new standards every few years, (e.g. 1966, 1985, 1999 and 2014) presents a changing goal post and is matched by a lack of consensus by experts on what constitutes validity evidence (Cook, 2014; Fromme et al., 2009). The most notable departure from the previous edition (1999) of the standards is the inclusion of a chapter on fairness in testing as an additional foundation standard alongside validity and reliability (American Educational Research Association et al., 2014). This aspect of test design has application to the content of OSCE stations, allowing for accommodation for students with disabilities and non-English speaking backgrounds; however, for the purposes of validation, ensuring accessibility in formats, and relevance in content for a broad range of contexts should be sufficient for the purpose of this study. Fairness, according to Schuwirth and colleagues, is ‘the level of accuracy to which decisions about candidates (e.g. good performance, in need of remediation, poor performance) can be made’ (2002, p. 927). Using these end-points and substituting the station-writer’s creative output, the OSCE station content and written instructions for the word ‘candidate’ implies that fairness as a validation criteria is appropriate when judging the effectiveness of the OWSAT tool. A valid tool should be able to reliably and fairly determine a grading for feedback on the OSCE station quality.

There are multiple approaches to gathering validity evidence described in the literature. Steps available to explore validity criteria, reported by Beckman and colleagues, include the use of Cronbach’s alpha and factor analysis to explore internal consistency, the use of test-retest to measure temporal stability, steps to investigate equivalence e.g. administering different forms of the same test, statistical measurements such as ANOVA looking at inter-rater reliability, and generalizability theory exploring reliability and measurement error (2004, p. 974). Furthermore, there was no particular level of evidence required before a tool is considered valid (Downing, 2003). Not all approaches to gathering validity evidence are suitable for inclusion in this study or in future research directions relating to this work.

Additional approaches to determine validity of a constructed tool potentially suitable for use in this research project include transparency regarding the steps involved in the creation and pre-testing, comparison with existing tools and the review by experts or peer groups using systematic approaches such as the Delphi technique (McKinley et al., 2008). Another interpretation of content validity asks whether the tool contains 'an appropriate sample of the items for the construct being measured' (Polit & Beck, 2006). Polit and Beck also detailed the use of ratings of item relevance by experts to reach agreement for inclusion of items in an instrument or tool (2006).

Design-based research, as an emerging method with unique challenges, requires a considered approach to construct validity evidence gathering. Cook documents the change in descriptive labels from validity types to validity evidence and explains the process as now presenting a hypothesis to be confirmed or discounted, requiring a well-constructed argument to present and speak to the evidence (2014). Polit and Beck (2006) advocate specifying the chosen validation method. Schuwirth and colleagues (2011) reporting on the consensus statement on Research in Medical Education from the 2010 Ottawa conference on assessment concluded the following;

> *'**Recommendation 2**: Developmental or design-based research should be realised through more than one single study, and be planned as a train of studies building the bridges between the idea, the pilot experiments, the improvements, the use in real life, etc.' (2011, p. 226).*

Support for this recommendation comes from the education literature. Kelly (2004) advises that meeting scientific claims for success is not enough in education design research. There is a practical approach and new criteria required that might need to be sourced from other disciplines, exploring, for example, whether an innovation is acceptable, efficient, efficacious, economical, and a true solution to a particular problem (Kelly, 2004).

Whilst many methods for validating a tool are unsuitable or beyond the scope of this research, some elements of validity evidence are relevant and warrant inclusion. Field-testing the tool with users other than myself to explore inter-rater reliability was not possible to achieve to the level required for validation, within the research time frame. Consequences, another construct validity category, require matching the use of the tool with the desired education outcomes. Further steps, such as these, to gather validity evidence will remain part of a longer-term goal should the tool meet other criteria such as utility and efficacy.

3.5 Design-based iterative process

The following three chapters will present the iterative process by which the education design-based research method was applied for this project. Specifics of the research process involved in the design of a tool (known as OSCE Writers Station Analysis Tool or OWSAT) for measuring OSCE errors in written station materials will be described in detail within these three chapters. Whilst traditional design-based research is undertaken in four stages, the fourth stage involves refinement post field-testing in the educational setting and was beyond the scope of this project. Hence, the ten core steps of this process are divided into three phases as demonstrated in Table 3-2.

Table 3-2: The three-phase approach to development of a tool to improve OSCE stations.

Phase One Tool Development	1	Recognition of problem and available resources for OSCE item writing.
	2	Literature review of assessment principles and key quality indicators.
	3	Review of post-OSCE debriefs for initial catalogue of errors.
	4	Thematic analysis of errors identified through literature review and available post-OSCE debrief reports.
	5	Creation of classification system for identified OSCE errors.
Phase Two Tool Refinement	6	Creation of OSCE item/station evaluation reviewer tool (OSCE Writers Station Analysis Tool or OWSAT).
	7	Presentation of tool at Canadian Conference in Medical Education in Quebec City.
	8	Workshop at ANZAHPE conference and Ottawa Conference on Assessment seeking additional feedback on tool.
Phase Three Tool Testing	9	Testing of tool against OSCE items in database to find potential additions to tool elements.
	10	Finalisation of tool including suitable graphics and formatting to improve utility.

Each step of this three-phase project will be described in detail incorporating an account of both methods and results. Given the iterative nature of design-based research, the many changes and reasons underlying subsequent versions of the tool will be outlined using chronological and reflective narrative. The method, data analysis and results for each phase will be provided in each separate chapter to provide an accurate and comprehensive account of the project outcomes.

3.6 Data and data analysis

Primary data for Phase Three was a collection of medical education OSCE stations; each station included the instructions for the simulated patient, the examiners, and the candidate as well as marking sheets, equipment lists and station objectives. The collection of OSCE stations were gathered from a number of sources, and provided examples from both undergraduate and vocational medical training in Australia, the United Kingdom and Canada. In order to maintain and enable efficient access to this collection, a database was created for this research. These stations were coded, identified by level of training and country, and reduced to a numerical label for the purpose of analysis. The process of categorising the errors encountered based on whether they would impact on validity, reliability, feasibility or educational impact was another step undertaken during the design and refinement of the tool.

A purposive sampling of the collated database of OSCE stations was used to explore whether the tool would identify potential errors for a station writer. This iterative process was undertaken to create, and then refine the language used for each question, trawl for previously unidentified errors and examine the utility of the tool. Stations were purposely selected to ensure an adequate mix of medical disciplines, level of training including pre-clinical and clinical undergraduate and vocational medical training examples, type of OSCE station (e.g. history, examination, diagnosis and management, procedural and communication) and mix of source institutions.

As will be described in Phase Three, the process of testing the tool against the database of OSCE stations was undertaken with the awareness of my personal ability to identify OSCE errors without the presence of the tool. This ability stemmed from an interest in the topic of the OSCE, and a scholarly approach to understanding current literature on the best approaches to assessing clinical competence. Use of this tool by novice OSCE writers when compared with those with high assessment literacy would have aided in qualifying its utility. However, the transferability of this tool to different users for field-testing is beyond the scope of this project. As will be described in Chapter 6, this project is predominantly limited to exploring the clarity, utility and comprehensiveness of the content of the tool. Further application of the tool is recommended for future studies.

3.7 Chapter conclusion

This chapter explored the background relating to the choice of method for this project, the ethical and practical considerations for the purposes of socially responsible research and described some of the processes undertaken during this thesis project. Design-based research offers a solution-based process for the purpose of responding to the concern about poorly written OSCE stations failing to be detected prior to examinations of clinical competence. An iterative process was undertaken to create, improve and test the tool to saturation whereby no further adjustments or additions to the tool were required. The following three chapters will detail the steps taken in the three phases of this project.

Chapter 4 – Phase One

4.1 Introduction

This chapter will describe the initial steps taken in pursuit of a solution for errors introduced into OSCE by station writers. The full description of the development of the tool for aiding OSCE writers or reviewers to improve the quality of OSCE was undertaken in three phases. Phase one of this education design-based research process included acknowledgement of the problem, exploration of the literature base for understanding the OSCE writing process and identifying specific OSCE station-level errors. Using the steps outlined Chapter 3, development of a solution began with delineation of the problem. These steps have been tabulated and are shown in Table 4-1.

Table 4-1: Phase One Tool Development

Phase One Tool Development	1	Recognition of problem and available resources for OSCE item writing.
	2	Literature review of assessment principles and key quality indicators.
	3	Review of post-OSCE debriefs for initial catalogue of errors.
	4	Thematic analysis of errors identified through literature review and available post-OSCE debrief reports.
	5	Creation of classification system for identified OSCE errors.

4.2 Step 1: The recognition of the problem and available resources for OSCE item writing.

The writing of an OSCE station is a creative activity where potential errors can occur during the OSCE development process. Errors in the writing phase for clinical examinations can originate from the wording of any component of an individual station. Created for the purpose of observing the student perform a task or tasks, each station requires documentation for all of the people involved in the proposed simulated interaction. The examiner, student, simulated patient (if one is involved) and staff setting up the station with equipment and other affordances need detailed instructions to ensure consistency in the way a station performs, particularly if there are two or more circuits of the same stations. Errors in the station writing can undermine standardisation of the students' experience. Flawed instructions or constructions in an individual station can cause misunderstandings between the expectations of the examiner or simulated patient, and the interpretation of the task by an examination candidate. Poor task clarity can then adversely affect the candidate's performance or the scoring of it by the observing examiner.

Quality improvement processes within the assessment of clinical competence using the OSCE format include many different interventions. These include planning and blueprinting meetings, standard setting meetings, reviewer feedback on the station components, piloting of stations, pre- and post-OSCE meetings for examiner and/or simulated patient training, and debriefing and psychometric analysis of station performances, all of which play a role in quality control in OSCE. All elements of a newly created OSCE station or even one which has been modified require close scrutiny, including instructions for the candidate, the simulated patient, examiners, the marking sheet and set up requirements for the room to ensure standardisation over multiple sites or circuits.

At a routine planning meeting for a summative OSCE in April 2009, a number of concerns about the design and content of several stations were identified. These issues were related to all stations, and were robustly debated by those with disparate views on best practice in OSCE station content. Decision-making required to complete the OSCE planning process was dependent on reaching a group consensus. Poor understanding of the impact of some of these decisions, when incorporated into the writing of a station, was identified during the meeting. A summary of the areas of concern and the potential impact on the examination or the students' learning is presented in Table 4-2.

Table 4-2: Summary of OSCE station planning meeting concerns (2009)

Issue raised	Potential effect on exam or student
Timing – too many tasks for station.	Students being trained to rush rather than reflect.
Timing – too much time for task duration.	Students might worry they have missed something. Examiners may turn waiting time into impromptu viva voce examination (i.e. a question and answer session between examiner and student).
Hand hygiene/ hand washing in OSCE station – before every station, - only for examination stations, or - not included in examination.	If not included, students will devalue this important component of patient safety. If only for examination stations – does not meet with patients' expectations. When student leaves examination station on way to next station e.g. history station, they are unlikely to stop to wash hands during reading time for next station. They may then shake hands with simulated patient in next station.
Content of history stations – some writers wanted to include systems questions from more than one system.	Insufficient time in one station for this and taking a comprehensive history. 'Mind reading' format requiring student to guess which system the examiner might want to hear. Training students not to think about which questions might be relevant to a particular patient presentation.
Some writers wanted the student to play the role of a GP for a particular station.	Authenticity issues. Teaching students to perform tasks outside their scope of practice sending wrong message through assessment.

The research aim for this thesis emerged from the desire to identify and eliminate these and other errors during the review process prior to an OSCE examination day. The hypothesis is that enabling OSCE station writers to become aware of potential defects in their creative output will improve the quality of the assessment of clinical competence. My observation of multiple errors found in stations that had been used in actual examinations was that writers appeared unaware of the inconsistencies or negative impact of the stations they had created. Making the invisible visible through a tool which linked good assessment criteria with the writing involved in the construction of a single OSCE station was considered a potential solution to this problem.

4.3 Step 2: Literature review of assessment principles and key quality indicators.

In Chapter 2, the key quality indicators of an OSCE were described. These indicators are validity, reliability, feasibility, authenticity, and educational impact (van der Vleuten & Schuwirth, 2005).

Exploration of available literature on best practice in the writing of OSCE stations provided few results, and none captured the situation of identification of pre-existing flaws prior to their implementation at pilot or in examination. Whilst analysis of the process of assessment using OSCE was a popular topic for publication, few, if any papers provided guidance, templates or checklists to assist with the item writing process. Assumptions of prior knowledge of psychometrics or assessment principles were prevalent in the available literature (Jansen et al., 1995; Newble & Swanson, 1988; Roberts et al., 2006; Sibbald & Regehr, 2003). Initial exploration of the literature uncovered only four resources, which outlined the practical aspects of conducting assessment using the OSCE format (Boursicot, 2003; Boursicot & Roberts, 2005; Curtis et al., 1994; Hurley, 2005). All of these touched on the topic of writing OSCE stations, but none specifically gave details to aid the OSCE writer to identify potential errors. Neither the video with an accompanying small booklet, created by the University of Toronto (Curtis et al., 1994), nor the book by Katrina Hurley (2005) provided more than scant detail on the aspect of OSCE station-writing, although many examples of written stations were provided in Hurley's book. The remaining two resources, written by Katharine Boursicot, including a journal article co-written with Trudie Roberts (2005) and an open-access PowerPoint presentation (Boursicot, 2003) contained some instructions for novice OSCE writers. Neither of these gave any indication of the complexity of the writing process for OSCE stations nor the degree of attention to detail, required to avoid creating stations with inbuilt errors. Since the initial literature review was conducted for phase one of this project it is recognised that there have been additional contributions to the body of knowledge on OSCE station writing, such as Zabar et al. (2013).

Criteria for good assessment have been well described within both medical and education literature (Boursicot et al., 2011; Chandratilake et al., 2010; Hattie et al., 1999; Norcini & Banda, 2011; Nulty et al., 2011; Schuwirth et al., 2002; van der Vleuten & Swanson, 1990). Hattie and colleagues (1999) also provided support for the concept of improving the measurement of performance through greater detail in the preparation of content specifications, limiting the domain covered within a station, and carefully weighting the scores from individual tasks being assessed during the performance. Papers describing principles of good assessment in clinical examinations include the Ottawa Consensus Paper on Criteria for Good Assessment (Norcini et al., 2011) and those by Chandratilake (2010) and Schuwirth and colleagues (2002).

Despite access to these and other publications describing considerations for quality in assessment, there is evidence as discussed here and within the literature review, that knowledge of good assessment criteria does not always result in valid and reliable OSCE stations being created (Auewarakul et al., 2005). The gap between best practice and the reality of flaws in assessment exists despite expanding volumes of publications on the subject. Attempts to bridge gaps between known research outcomes and implementation into practice has resulted in an expanding body of work in the field of educational research (Akkerman et al., 2013).

Not all examinations are subjected to all of the layers of possible quality improvement activity, largely due to lack of resources. Sometimes processes are omitted due to a lack of understanding of the benefits or methodology of these practices. For example, piloting of stations is a particularly expensive quality improvement process, and psychometric analysis requires an understanding of statistical principles or access to a statistician that may not be possible for all institutions. Insufficient numbers of faculty with a willingness to contribute to a reviewing process of the wording of the OSCE stations, or those lacking in assessment literacy in this area might limit access to this less expensive option for improving station clarity or feasibility.

Popham (2011) argues that deficiencies in assessment literacy contribute to poor quality in educational measurement. As seen within the literature review in chapter 2, the concept of assessment literacy provided an explanation for the disconnect observed in the OSCE setting, between both proposed and implemented OSCE stations and the wealth of literature available on best practices in the assessment of clinical competence. Given the aim of this project was to aid station developers and reviewers to identify station-level errors, assistance with translating the education literature into an accessible format through the application of a suitable tool may aid OSCE writers or reviewers to improve overall assessment literacy as they are encouraged to incorporate best practice into their clinical assessments.

Broadening the literature review to include German language materials led to the identification of a pre-existing tool to assist OSCE station writers to review their items for quality improvement (Schultz et al., 2008). A team at Heidelberg University working to improve the quality of the combined assessment of clinical training by the universities of Germany created a tool to be used by the OSCE station reviewers (Schultz et al., 2008). Following the introduction of their tool, positive improvements in the reliability of their OSCEs were observed over three years. In particular, the reliability or internal consistency of their examinations demonstrated an improvement from a Cronbach's alpha of 0.50 to greater than 0.8 over three years (2008, p. 669 English translation of German by K.Brotchie). Use of the Heidelberg tool was not an isolated intervention reported in their study, as other measures were simultaneously reported which combined to produce these results (Schultz et al., 2008, p. 668). The station design review tool was used in conjunction with examiner training and feedback of psychometric data to the station designer. All three elements were considered to be responsible for this dramatic improvement in their clinical examination and it is therefore not possible to ascertain the true impact of an individual intervention in the assessment process.

As discussed in chapter 2, the level of assessment literacy an individual possesses has relevance when undertaking the role of OSCE station developer or reviewer. The tool from the University of Heidelberg medical school (Schultz et al., 2008) was quite simple in its structure, but assumed a significant degree of assessment literacy, which may limit the utility of this tool in other settings. The aim of this thesis is to create a structured evaluation tool with broad reach, to aid all OSCE writers or reviewers, not just novices, but also those with more extensive knowledge and experience to identify errors. All aspects of the OSCE writing process, including instructions for candidate, standardised patient, examiner and instructions for the set up the equipment for the examination need to be considered for potential flaws. Recognition that an OSCE reviewer assistance tool already exists does not detract from the need to create one with broader utility across the health professions, to meet the needs of those with varying degrees of assessment literacy skills.

4.4 Step 3: Review of post-OSCE debriefs for initial catalogue of errors.

Improving the OSCE quality requires a commitment to audit and reflection processes to appraise station performance. Post-OSCE debriefing, involving input of examiners, simulated patients and OSCE support staff following an examination is a qualitative evaluative process that can contribute over time to improved assessment (Sudan et al., 2015). Feedback from OSCE candidates is also valuable in understanding the impact of components of the examination including, for example, alignment of candidate instructions and marking guides, providing useful insight into desired improvements in station design for the item writer (Chipman et al., 2007). These opportunities are not universally adopted by all institutions running the OSCE and post-OSCE debriefs, and may be more achievable with smaller candidate numbers resulting in a shorter examination day. Other quality improvement initiatives include pre-OSCE station standard-setting processes, where errors in OSCE stations can be identified prior to the actual examinations and video review of stations and student performances in the actual examination using internal and external reviewers (Barry et al., 2013). The use of the whole of faculty at a small rural medical school for review of OSCE station wording pre-examination review was the basis of a presentation at the Ottawa conference on assessment in Kuala Lumpur in 2012 (Brotchie et al., 2012).

Field notes recording examiner and simulated patient feedback during a routine post-OSCE debrief meeting following a summative OSCE form the basis of the following analysis document. Written from an oral transcript of the meeting held at the conclusion of one of the year-level multi-site examinations at one of the institutions involved in this research, the variety of errors and annoyances discussed during one exam debrief meeting provided rich initial data for analysis of the types of errors occurring in OSCE stations. Extensive feedback was supplied by up to thirty examiners at the meeting, who identified elements of concern in eight of the ten stations, as presented in the first three columns of Table 4-3.

Table 4-3: Analysis of Post-OSCE Debrief 2011

Station Number	Type of Station	Issues reported by examiners and simulated patients	Type of Error(s)
One and Two	History/Diagnosis	No issues reported	-
Three	History / Pathology / Diagnosis / Management	Issues with congruence between patient history and content of checklist '*Station needed rewording.*'	Congruence between different elements e.g. patient history information and marking sheet checklist. Timing error.
Four	History and Examination	Issues with candidate instructions, and structure of tasks. No marks for hand washing?	Candidate instruction errors. Educational impact.
Five	Examination / Report findings / Read an X-ray	Issues with structure of station – too many tasks for the time. Quality of x-ray poor, difficult to interpret.	Timing error. Errors in Props.
Six	Examination/ diagnosis	Too much time allocated for tasks. Task not taught at this year level? Hand washing not included in marking sheet criteria.	Timing error. Curricular relevance error. Educational impact error.
Seven	Procedural station	Station too long. Student instructions unclear, regarding context and examiner instructions.	Timing error. Clarity of task for student instructions error. Context issues. Examiner instructions unclear.
Eight	History/ECG / Diagnosis	Issues with props and checklist items repeated.	Prop error. Marking guide error.
Nine	Pathology / Counselling	*'Students wanted to read the whole report so* ***one examiner*** *just pointed to the bottom. Students assumed that basic bloods like an FBE would have been done already if they made it to theatre. Bone marrow biopsy wasn't in checklist but FBE and LFTs/ CT were there. One examiner* ***prompted*** *for basic tests…'* *'Disaster of a station'* (See appendix B).	Issues with structure of station and appropriateness of items in checklist.
Ten	History / pharmacology explanation	Poorly designed station. Inappropriate context, station content didn't match good clinical guidelines for management. No name for patient.	Issues with structure of station –Issues with clinical management details. Doesn't meet current best clinical practice. Patient safety issues.

More details regarding concerns about the OSCE station performances presented in (Table 4-3) are provided in Appendix B Transcript of early clinical years post-OSCE debrief.

No aspect of the OSCE process was absent from the critical feedback provided with concerns relating to all components of the station-level instructions. These flaws were identified within the student instructions, notes about the role of the simulated patient, instructions to the examiner and the marking sheet. Station props were also criticised, for example the use of a facsimile of a real electrocardiogram that was reproduced in such a format that it was too small to be easily interpreted and produced by an ECG machine that students would not have encountered during their clinical placements.

The effect on the students of these errors is difficult to quantify, as is the effect on examiner retention following a day spent in a flawed station observing students struggle to perform in a high-stakes environment. Many of the errors are likely to have prevented students from demonstrating their ability to meet the objectives of the station, even if they were able to perform the task in a normal clinical situation. In addition, errors were identified that rewarded students for restating information provided in the stem (students' instructions), and some concerns were expressed about the accuracy of the clinical content undermining students with up-to-date clinical knowledge where the station rewarded the use of out-dated clinical guidelines. Examiners reported altering scripts and marking guides to try and assist the students, further undermining both reliability and validity of the summative assessment, where these changes could affect some but not all students in the cohort.

4.5 Step 4: Analysis of errors identified through a literature review and post-OSCE debrief reports

Identification of the elements to be considered for inclusion in the tool began with reviewing the feedback reports from examiners and simulated patients at this and other post-OSCE debriefings following the running of examinations across different year levels. Failure to recognise these flaws in the lead up to the OSCE posed significant challenges for students in demonstrating their level of skill during that examination. The impact of these design flaws was not observed subsequently to have resulted in any planned amendments for subsequent use of these flawed stations. This meant a high likelihood that these faults would persist through to implementation in future summative OSCEs at that year level.

Several options existed to categorise the errors identified in just one of the OSCE debrief reports. The final column of Table 4-3 provides examples of the types of errors found within the post-examination reflections, including timing errors, task communication errors, equipment or prop errors, structure or sequencing errors, currency or evidence-based errors, and patient-safety or educational impact errors. Alternative categorisation could follow the effect of the errors on the quality of the assessment. Errors affecting validity, reliability, feasibility, authenticity and educational impact were all recognised within the one post-OSCE debrief. Yet another option is to categorise based on which component of the paperwork for the station holds the key to the flaw. Errors, from a station-writer's perspective, could emerge from the composition of station, derived from the objectives of the station being too ambitious, or within the directions provided for candidate, simulated patient, examiner or OSCE technical support staff.

4.6 Step 5: Creation of classification system for identified OSCE errors

Raising awareness of the concept of errors in OSCE stations created during the writing phase of the assessment development process was the theme that inspired the abstracts submitted to medical education conferences during this research project.   This journey commenced with the Ottawa conference in Kuala Lumpur and a presentation on the whole-of-faculty contribution to OSCE station reviews prior to the examination as a quality improvement process (Brotchie et al., 2012). Subsequent conference abstracts explored the various stages of the development of the tool beginning with the initial OSCE error classification presentation at the Asia Pacific Medical Conference in Singapore in January 2013.

Analysis of the key concerns arising from post-examination debriefing reports from multiple undergraduate year level OSCEs formed the foundation for a presentation on a classification system of errors (Brotchie et al., 2013). Initially limited to 'sins of omission' and 'sins of commission', i.e. not doing things that ought to be done and doing things that ought not be done, the refinement of the list of errors evolved into a list of ten tips for writing OSCE items. Whilst not exactly being about unhelpful OSCE writing, these tips covered the areas of potential errors previously identified through examiner or simulated patient debrief reports or through the literature review. This ten-question list aimed to assist an OSCE station writer or reviewer to reflect on the creative aspects of the content and process outlined in the newly created or recycled station document as part of the quality improvement pre-exam processes. This list, as presented at the Asia Pacific Medical Education Conference in Singapore in 2013, is found in Table 4-4.

Table 4-4: List of ten tips for writing OSCE stations as presented in Singapore in 2013

a) Is there a better tool for the job?

b) Do you have the correct resources?

c) Is the timing okay?

d) Is the sequencing of tasks okay?

e) Does the station follow best clinical practices?

f) Is the task clear and allows candidates to perform what is being graded?

g) Does the checklist or rating scale match the objectives?

h) Is it safe for the simulated patient to be in this station?

i) Does it all match the current curriculum for this year level?

j) Are we negatively influencing our students through this station?

The audience in Singapore who attended my end of session presentation responded enthusiastically to the concept of assistance with improving OSCE station writing. The large auditorium at the National University of Singapore Cultural Centre contained less than one hundred people, the most notable being Professor Ronald Harden, who published the seminal work on OSCE (Harden et al., 1975). Naïve to the process, I can no longer recall the content of the three questions that followed my presentation. At the conclusion of the session I was amazed to see a group of predominantly female, Asian conference attendees, waiting to talk with me as I left the stage. All were keen to know more about prevention of errors in OSCE. It was clear from our discussions that my experience in observing flawed assessments and wanting to find a solution was not unique.

4.7 Chapter conclusion

This chapter explored the first phase of the iterative design based research approach to the development of a tool to assist with improving the quality of OSCE station writing. Critical reflections by examiners at a post-OSCE debrief session provided insight about the types of errors observed during the examination and formed the basis of an oral presentation at the Asia Pacific Medical Education Conference in January 2013 (Singapore). Recognition of the multiple errors per station enabled the creation of a classification system for errors in OSCE writing.

The steps taken along the process from recognising the problem this thesis intended to address, through to the creation of a list of questions to explore when writing or reviewing OSCE stations were outlined in the first five steps of this three-phase project. The results obtained during each step of this phase of the research journey were described along with examples to illustrate the contributions to the design process. The following two chapters will detail the steps taken in the subsequent phases of this project beginning with phase two and the creation of OWSAT (the OSCE Writers Station Analysis Tool).

Chapter 5 – Phase Two

5.1 Introduction

This chapter will describe the next steps taken in the iterative three-phase process in the creation of a tool for improving the writing of OSCE stations. Phase one concluded with a simple list of questions to assist with uncovering or preventing flaws in OSCE stations. Phase two of this education design research project progressed from the list of errors to the creation of the tool labelled OSCE Writers Station Analysis Tool (OWSAT). The tool was then presented for feedback at an international peer reviewed conference in medical education, and then subsequently at two workshops, one locally and one internationally for peer reviewed commentary. The main steps in phase two are outlined in Table 5-1.

Table 5-1: Phase Two Tool Refinement

Phase Two - Tool Refinement	6	Creation of OSCE item/station evaluation reviewer tool (OSCE Writers Station Analysis Tool or OWSAT).
	7	Presentation of tool at Canadian Conference in Medical Education in Quebec City.
	8	Workshop at ANZAHPE conference and Ottawa Conference on Assessment seeking additional feedback on tool.

5.2 Step 6: Creation of OSCE Writers Station Analysis Tool (OWSAT)

Following the presentation in Singapore of the ten tips for OSCE writing, the next iterative process was to transform these items into a tool for writers and reviewers. The creation of the OSCE Writers Station Analysis Tool (OWSAT) involved an iterative process. The list of errors required transformation into a format that encouraged critical reflection of stations as described on paper, without seeing how they performed in situ. The supposition was that any tool triggering insight for OSCE reviewers could help OSCE writers during the writing process. Exploring each suggestion from the Singapore presentation and rewording as required led to the first draft of the OWSAT tool. Naming the tool occurred later in the iterative process; however, I will use this title prematurely in the interests of writing brevity whilst maintaining clarity.

Applying the Socratic method to the basic error list identified three key questions relating to concerns about how the tool would function for the purpose it was intended. Awareness of issues of clarity, format and comprehensiveness or inclusiveness resulted from contemplation of how a novice or experienced OSCE station writer or reviewer would interpret the statements as they presented in the initial presentation. The iterative process was necessary to work on these components of clarity, format and inclusiveness. Each of these three aspects will be discussed in relation to the evolution from a list of errors to the emergence of the first iteration of OWSAT.

5.2.1 Tool Clarity

A series of questions with the potential to direct OSCE station reviewers to identify potential flaws was required; however there were issues with the clarity of the initial list of errors in OSCE. An attempt at classification or taxonomy for the construction of the list was undertaken through the grouping together and merging of questions relating to errors of a similar domain. This simplification had the consequence of reducing the intelligibility of the communication. Combining more than one potential flaw of a similar domain within a question in OWSAT was potentially misleading. Consequently writers may miss some errors because the directions were not direct enough or lacked a suitable rubric. Instead of reducing the number of error-identification questions, working to improve comprehension of each probing question listed for inclusion in OWSAT was undertaken for the purposes of tool clarity.

An example of concern relating to clarity is found in the first question: **'Is there a better tool for the job?'** For those with a good understanding of assessment vocabulary the proper term is 'instrument' rather than tool. It is unlikely that many of those involved in OSCE writing would be troubled by assessment semantics. This question was intended to highlight errors in the choice of overall assessment instrument, or, the use of the instrument to perform a task that could be undertaken by a more cost-effective assessment instrument. Using the time in a clinical station to ask basic knowledge questions is the commonest example of an inappropriate choice of assessment instrument. Given the cost involved in running a clinical examination, using OSCE as an assessment of lower order questions is a poor economic choice (Schuwirth & van der Vleuten, 2003). However, the initial wording of the first OWSAT question may not facilitate awareness of this distinction in a novice OSCE station writer.

The ambiguity of the first statement could also mislead the station writer or reviewer to consider the use of resources rather than the choice of assessment format for each task. Directing attention from the overall assessment, the language used in this question could lead the tool user to question the use of props, simulated patients, consumable materials or equipment used within the station. This important consideration in the OSCE station development process is the subject of the second question in the list: '**Do you have the correct resources?'** This question also lacked immediate clarity given that the concept of resources may not be obvious to those who have not been involved in the actual setting up of stations or recruitment of simulated patients for the roles depicted in an OSCE station. Removal of potential areas of confusion required definition of terminology or some other solution to facilitate adoption of a shared understanding of OSCE related language.

Multiple options were explored for the purpose of improving clarity, beginning with refining the wording of each question. The majority of attempts to illustrate or define terms or concepts resulted in an increased word count for each question or the provision of additional text in the tool list. The risk posed in elaborating for the purposes of comprehension was sacrificing the brevity required for the utility of the tool.

5.2.2 Tool Format

Apart from refining the list of questions to direct a critical exploration of station wording, discussion now evolved around potential formats for the tool and the evaluation key during this early iteration. The stage of transformation from a list of basic errors to becoming a functional tool coincided with an exploration of possible formatting options for the tool. In particular, many options were identified in the search for the best possible structure for the answer code. Use of dichotomous or tri-partite solutions was debated, along with the potential for a Likert scale to be used for grading of the errors. Although most of the questions are of a closed or binary format where only yes or no are possibilities, for others, a more complex answer may require the option of provision of comments to direct the writer, post review, to the components requiring change.

Consideration of potential formatting options for the tool ignited some brainstorming, including possibilities such as creating a mobile device application. Limitations with respect to available skills, budget and time resulted in the reality check around feasible options within the confines of this Master's project. The option of a web-based application was considered the best option for global distribution of the concept of an aid for OSCE writers. A paper-based and published version within medical education literature was another potential distribution mechanism with broad reach, yet with fewer resource requirements in the short term. Providing both a paper-based version of the tool, and exploring the use online software programs that could be accessed via email or web links, were the preferred options for the tool format for the purpose of this project. The use of technology to create an accessible method for dissemination of OWSAT to OSCE station writers and reviewers was appealing but impractical in the time frame between the presentation in Singapore and the subsequent launch of the still unnamed tool to an audience at a peer-reviewed international conference in medical education in April 2013. A paper-based tool was the only feasible option during this period.

5.2.3 Tool Inclusiveness

Variation in the level of expertise in OSCE writing or reviewing and degree of assessment literacy presented a challenge for achieving a broad utility for the tool. Creating an additional step in the tool was contemplated, to aid with assisting less experienced reviewers or writers, or those unfamiliar with the preferred OSCE terminology in my context, to understand the purpose of each element encountered.

An ambitious goal for the long-term direction of OWSAT was to improve the quality of OSCE stations globally, an objective that requires a high level of acceptance for both the format and the perceived effectiveness of the error detection tool. The most daunting step envisaged in this process would be convincing self-anointed experts in the field to apply the tool, given the lack of attention in the literature to the possibility of station-level flaws undermining validity of published studies using the OSCE format. Confirming that the simplicity of the tool wording did not appear to be patronising this group of experienced medical educators was essential. The creation of the tool to facilitate communication between novice medical educators and their more experienced colleagues over the existence of flaws within OSCE station-level development review meetings using a structured approach to highlight areas of concern was a driver for this project. Given the observed frustration and lack of response to previous examples of written feedback on poorly performing stations both pre- and post-OSCE experienced at more than one institution, this created an imperative for maximising tool utility. Ensuring inclusiveness through enhanced clarity to improve tool item comprehension required the adoption of multiple format options including both paper-based and online tool versions.

The use of technology to assist with tool utility, for example producing a video or web based explanatory note, was considered one option for maximising understanding about the purpose and process of using OWSAT. Creating a web-based application for the OSCE error detection tool questionnaire was another option identified. In this setting, a hyperlinked prompt over a question could lead to additional explanatory notes being revealed. Having a second page with explanatory notes as a FAQ (frequently asked questions) sheet, or generating a paper based user guide to the tool were other solutions proposed for the purpose of tool interpretation. Some options, including a one-pager tool explanation were immediately accessible whilst others could be undertaken as further development beyond the term of this project.

5.3 Step 7: Presentation of tool at the Canadian Conference in Medical Education in Quebec City.

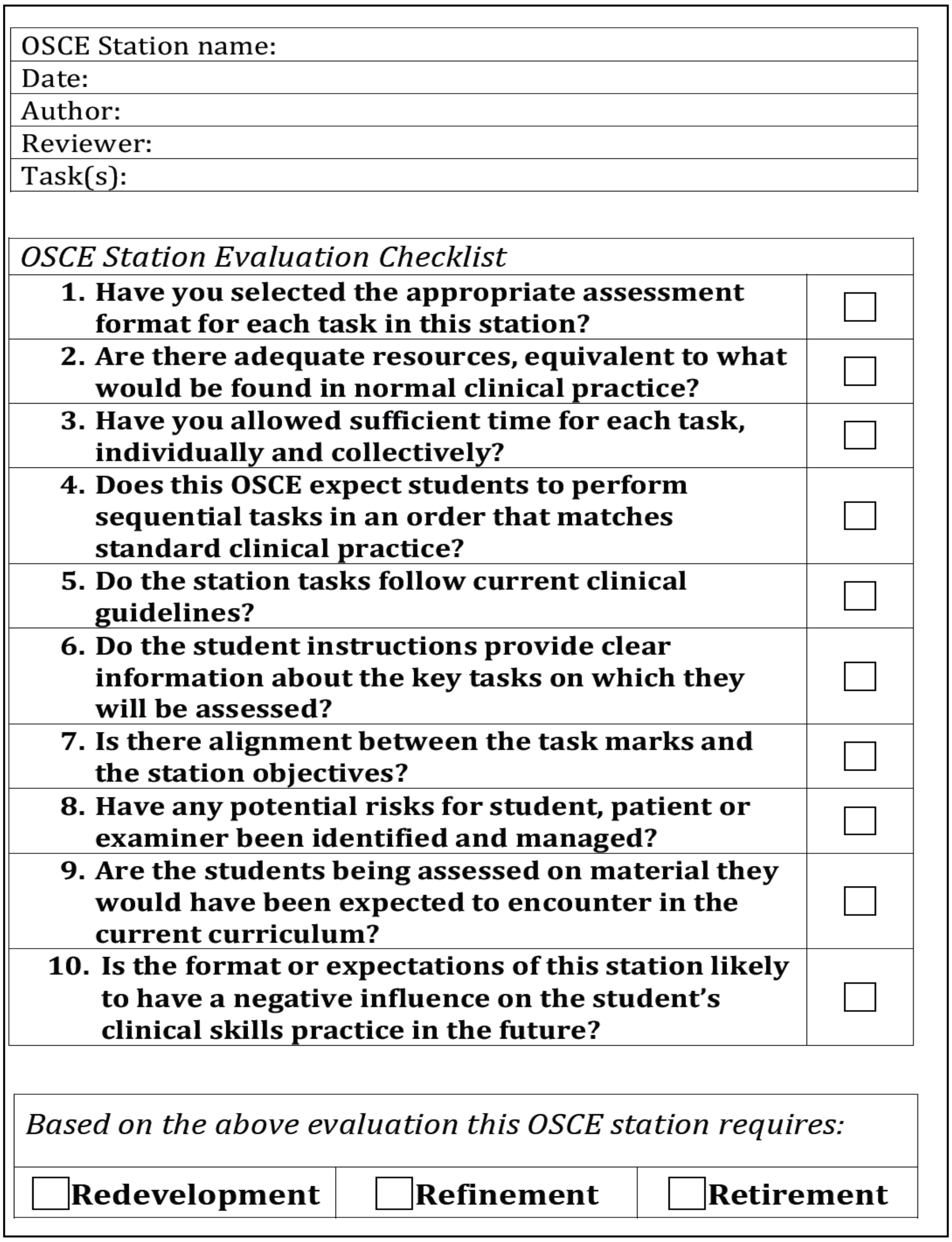
Skills development in research and scholarship includes the writing of abstracts for potential peer-reviewed conference presentations at national and international discipline related scientific meetings. The submission of abstracts following a sequential anticipated trajectory for the development and refinement phases of this research project was unexpectedly successful, supporting a hypothesis of OSCE station-level flaws forming a significant, yet neglected topic in medical education and assessment. This abstract acceptance success led to the addition of immovable deadlines at the point of conference presentations for the next few steps of the tool development process as well as opportunities for feedback and peer input into the different phases of the project.

Transformation from the original list of ten questions exploring OSCE station-level errors, presented at APMEC in Singapore in January 2013, into the first iteration of the OSCE Writers Station Analysis Tool (OWSAT), required attention to the underlying language. Initially there was no increase to the number of questions from the original list, when transformed into the paper-based OWSAT; however, OWSAT provided more specific instructions to direct the reviewer to explore elements relating to each of the ten domains.

Enhanced question phrasing directed the reviewer to examine specific components of stations including the format, resource lists, task description and allocated timing. The modification process improved the clarity of instruction and inserted a dichotomous response option for each question. A global decision tree at the conclusion of the OSCE evaluation process allowed a self-directed interpretation of the responses given during the application of the tool to a particular OSCE station. Options included redevelopment, refinement, or retirement, and this determination was related to the number or seriousness of problems or errors detected through the use of the tool and the proposed outcome for that station.

The first iteration of the OWSAT was formally presented in Quebec City at the Canadian Conference in Medical Education in April 2013. A central theme of this conference related to patient safety and errors in clinical practice providing a useful linkage with the message of my presentation relating to errors in the assessment of clinical competence. Presenting the background to the tool development and introducing OWSAT to the audience, also created an opportunity to educate on different types of flaws and their effect on various aspects of the OSCE process. The first iteration of the tool (OWSAT) is shown in Figure 5-1.

Figure 5-1: First iteration of OWSAT tool as presented in Quebec City in 2013



OSCE Station name:
Date:
Author:
Reviewer:
Task(s):

OSCE Station Evaluation Checklist	
1. Have you selected the appropriate assessment format for each task in this station?	□
2. Are there adequate resources, equivalent to what would be found in normal clinical practice?	□
3. Have you allowed sufficient time for each task, individually and collectively?	□
4. Does this OSCE expect students to perform sequential tasks in an order that matches standard clinical practice?	□
5. Do the station tasks follow current clinical guidelines?	□
6. Do the student instructions provide clear information about the key tasks on which they will be assessed?	□
7. Is there alignment between the task marks and the station objectives?	□
8. Have any potential risks for student, patient or examiner been identified and managed?	□
9. Are the students being assessed on material they would have been expected to encounter in the current curriculum?	□
10. Is the format or expectations of this station likely to have a negative influence on the student's clinical skills practice in the future?	□

Based on the above evaluation this OSCE station requires:

□**Redevelopment** □**Refinement** □**Retirement**

Questions following the oral presentation gave insight into the acceptability of this tool and included a request for online availability of OWSAT. A presenter from the previous day had researched quantitative error identification processes and discussion relating to merging the qualitative aspects of my research with the quantitative psychometric analysis provided an opportunity to reinforce the message relating to detecting avoidable errors during the OSCE development stage. The request for availability on-line during the post-presentation question period indicated acceptability of both the tool and the future adaptation of OWSAT to an on-line format.

Each of the questions posed in OWSAT potentially link to problems with reliability, feasibility and/or educational impact. Work undertaken for the presentation in Ottawa involved expanding on the classification of errors types in OSCE stations presented in Singapore in January 2013 (Brotchie et al.). Renewed attention focused on this problem produced two key elements suitable for incorporating into a future version of the tool. A systematic approach to thinking about errors linked with an understanding of where the error could impact on the quality of an assessment emerged from the work preceding this conference presentation. For each question designed to pick up errors in a different component of the OSCE development process it was possible to describe whether it would have an effect on any of the aspects of the utility of assessment model described by van der Vleuten in 1996. As outlined in the literature review, the five facets of his model include reliability, validity, educational impact, acceptability and cost.

Included in the presentation for the launch of OWSAT in Quebec at the Canadian Conference in Medical Education was an explanation key for each question and the linkages between the question domains and the potential locus of effect of a particular error. These additions are shown in Table 5-2.

Table 5-2: Relationship between error and possible adverse effect on station performance

Questions	Possible Error Identified
1. Have you selected the appropriate assessment format for each task in this station? (Miller, 1990)	
Is it a straight knowledge test? (Use a written examination – do not waste your resources) Is it vital that student can perform this task (pass/fail) (Make it a hurdle in the course, not an OSCE station) Is it something better observed over time? (Use In training assessment)	Feasibility Educational impact
2. Are there adequate resources, equivalent to what would be found in normal clinical practice?	
Do you provide gloves? All hand sizes? Do you have enough cotton wool pieces, paper clips so no one is reusing even with the simulated patient? A step stool for every circuit and every examination couch to ensure safety for the short students?	Feasibility Reliability Educational impact
3. Have you allowed sufficient time for each task, individually and collectively?	
What message do you send when you have too many tasks for the student to demonstrate competency in within the one station? BEACH data – 8 minutes is average time for GP consultation to perform history, examination, diagnosis, management and patient education. Is this also appropriate for an undergraduate student?	Feasibility Reliability Educational impact
4. Does this OSCE expect students to perform sequential tasks in an order that matches standard clinical practice?	
Expecting the student to talk about ordering basic tests in a post-operative patient, who is likely to have had these done pre-operatively History before examination, before investigation. Consider not ordering any tests?	Reliability Educational impact
5. Do the station tasks follow current clinical guidelines?	
If the students know what is current and the station writer does not make sure they are up to date, the good student will be disadvantaged. Particularly relevant for ACLS and BLS – which also need people to be clear about what is in hospital and what occurs outside hospital	Reliability Educational impact

Questions	Possible Error Identified
6. Do the student instructions provide clear information about the key tasks on which they will be assessed?	
What is the patient’s name? What are you expecting the student to perform Would any clinician stepping into the room know what you want them to do without the provision of prior coaching?	Reliability
7. Is there alignment between the task marks and the station objectives?	
Are you providing marks for tasks a student would not know to do? What about the weighting of marks? Do marks reward skill and time use wisely to enable discrimination between good and poor learners? Do you let the examiners know what the marks are for various elements? (area where you might encourage examiners behaving badly)	Reliability
8. Have any potential risks for student, patient or examiner been identified and managed?	
Use of sharps, gloves, ophthalmoscopes, otoscopes, etc. Reflexes performed with patient sitting on edge of portable couches, Kneeling on couch for ankle reflex (thyroid exam) Step stools for shorter students Height of instructions to candidate on door – too low or too high for some	Feasibility Reliability Educational impact
9. Are the students being assessed on material they would have been expected to encounter in the current curriculum?	
Includes scope of practice concerns, e.g. Candidate instructions stating ‘You are a GP’ for a first year medical student. Review learning objectives and curriculum content to align with assessment content - ideally this takes place during the assessment blueprinting process.	Reliability Educational impact (may have a positive or negative impact)
10. Is the format or expectations of this station likely to have a negative influence on the student’s clinical skills practice in the future?	
No gloves provided in the station. Reason provided - there is not enough time in 8 minutes to suture and put on gloves What message are you sending the students? When it gets too busy forget PPE (personal protective equipment) What is the educational impact on the examiners as well as the students? Medial collateral ligament strain – orthopedic referral or CT/MRI in the examiners instructions.	Educational impact

5.4 Step 8: Workshop at ANZAHPE conference and Ottawa Conference on Assessment seeking additional feedback on tool.

Following the launch of OWSAT in Canada, abstracts accepted for two medical education conferences provided an opportunity to hold workshops for the purpose of obtaining peer feedback on the utility and comprehensiveness of the tool. The first of these was in the format of a PeARL (Personally Arranged Learning Session) at the Australian and New Zealand Association for Health Professional Educators (ANZAHPE) conference in Melbourne June 2013. The second feedback session was held in Ottawa in March 2014 at the Ottawa Conference on Assessment, and was conducted as a peer reviewed workshop. A minor modification to the format of the tool was made between the Canadian Conference in Medical Education presentation in 2013 and the ANZAHPE PeARL session, with the addition of a column for marking yes or no as a response to the questions.

PeARLs are sessions where the presenters have only five minutes to present the topic and outline their aspirations for the ensuing open discussion. In this unique format, there is an opportunity for novice researchers to tap into the wisdom of more experienced medical educators, as well as creating a platform for dissemination of research interests to the audience and conference attendees. The PeARL conducted in Melbourne was of a similar format to the subsequent workshop held during the international conference on assessment in Ottawa in 2014, and both workshops provided valuable insights into the clarity and utility of various components of OWSAT.

The opening segment of both sessions provided participants with an opportunity to add to the catalogue of errors with their own experiences of OSCE stations that failed to perform at an acceptable standard. On reflection, the entertaining narration of OSCE errors and aspects of the theatrical elements of simulated performance-based assessment did not reveal any deficits in the tool as it was presented. A unique addition to the list of possible flaws in resources or format choice was provided with the narration of an OSCE station on certification of death where the moulaged simulated patient fell asleep during the long day and was observed to be snoring whilst portraying a corpse.

Based on the first component of the workshop, seeking gaps in the tool from the volunteering of experience with flawed OSCE stations from the workshop audience, no additions to the tool were contemplated. All errors presented by the audiences in these sessions could have been identified within the available questions in OWSAT. With the exception of the snoring corpse, had the tool been used for station evaluation, errors should have been detected. The presence of suitable domains for reviewing the use of a simulated patient in a death certification station, including choice of the right resources and simulated patient safety from a psychological perspective is unlikely to have led to a different outcome without sufficient imagination with respect to the outcome of playing a role lying very still for an entire examination day. Input from simulated patients or trainers to the pre-OSCE quality improvement meeting may add to the effectiveness of both the meeting and the tool effectiveness.

Workshop participants were then provided with an OSCE station with extensive flaws purposely created for the session. An opportunity to identify as many errors as possible was provided prior to the introduction to the OWSAT. Approximately five errors were identified by the majority of the audience in Melbourne prior to the tool being applied. A subsequent review of the station using the tool was conducted and feedback sought on the usefulness of the tool for this purpose. The ability to detect more errors after the tool was applied surprised audience participants and discussion ensued regarding the impact of the tool on the OSCE as an assessment format.

The same format was followed for the second workshop in Canada nine months later, where a smaller audience of international delegates included national licensing station developers and novice medical educators. There was hesitant sharing of OSCE experience, an enthusiastic hunt for errors within the provided flawed OSCE station and a robust discussion following the introduction of the tool to aid with error identification. Additional recommendations for inclusion in the tool following the workshop included a question relating specifically to the presence of cultural, ethnic or gender bias within the OSCE.

5.5 Chapter conclusion

This chapter explored the second phase of the project, the creation and early field-testing of the OSCE Writers Station Analysis Tool (OWSAT). Using an iterative approach through design-based research, phase two of this project transformed a list of questions regarding potential OSCE station-writing errors to the creation of the tool. International and national presentations formed key steps in the exploration of the feasibility and possible utility of the tool. The following chapter will detail the final steps taken in the third and final phase of this Master's project.

Chapter 6 – Phase Three

6.1 Introduction

This chapter will provide the final steps undertaken in the creation and refinement of the design-based research process leading to the development of the tool labelled OSCE Writers Station Analysis Tool (OWSAT). Phase one described the background and context providing the foundation work for the tool creation. Phase two of this design-based research progressed from the list of errors to the creation of OWSAT. Phase two concluded with a series of workshops testing the tool and gathering peer feedback. Phase three contains the conclusion of this iterative process whereby the tool was tested against a database of OSCE stations from multiple sources. The two steps of phase three in the tool creation process are outlined in table 6-1.

Table 6-1: Phase Three – Tool Testing

Phase Three - Tool Testing	9	Testing of tool against OSCE items in database to find potential additions to tool elements.
	10	Finalisation of tool including suitable graphics and formatting to improve utility.

6.2 Step 9: Testing of tool against OSCE items in database to find potential additions to tool elements.

Over the years I have developed a database of over 500 OSCE stations. Initially the database of OSCE stations was designed to support my role as a station writer. An awareness of my interest in the topic of OSCE errors resulted in additional collections of existing stations being donated to me by academics with ties to other institutions. Stations were added to the database through my roles at several institutions, covering both undergraduate and postgraduate level examinations. Some stations were acquired through participant handouts and OSCE writing workshops that I had attended at national and international medical education conferences. Furthermore, the literature review provided some examples of working OSCE stations, predominantly located within the print media. Incomplete stations were also added to the database, as sources of errors were visible within the documentation, adding to the overall picture. The sources of the OSCE stations included institutions in North America, United Kingdom/Europe, and representative examples from four different states in Australia, all of which were used in the testing phase of this design-based research project. The collection of OSCE stations commenced in late 2008 and covers the period up to April 2015.

The OSCE stations were in both paper-based and electronic formats. The total number of stations in the database numbered over 500. Not all stations were complete; some were missing resource lists, others the explicit candidate instructions, yet all were suitable for the process of testing the OWSAT tool. In the incomplete OSCE materials, information about the candidate instructions was included within the examiners' instructions making this aspect of the station clear for the purposes of tool testing. Some duplicates of the electronic version were present in the paper-based stations collection, and multiple drafts were examined in both formats as these contained valuable information regarding previously detected errors. The stations were de-identified using an institutional and year level code to ensure anonymity within the testing phase of the tool.

The first step in this phase included conversion of the OWSAT tool from a one-page list of questions on possible sources of errors to a more comprehensive on-line version using an online software program. Qualtrics®, a web-based software program for survey creation, or Survey Monkey®, another popular survey instrument, were recognised as potential platforms for the on-line version of the tool. This was seen as a valuable step, as an online system could be accessed via smart mobile devices and would make it easily accessible in a non-paper based culture. Qualtrics® was a platform provided for our University-run OSCE, where iPads loaded with the assessment marking sheets and results are fed back to a central server. Qualtrics® was identified as meeting the requirements for the OWSAT tool with the potential of varied platforms for delivery. It was intended that responses to the OWSAT could be sent to the user at the conclusion of the Qualtrics®-based survey to aid with the next step in the OSCE improvement process. Qualified opinions on whether major or minor revisions or abandonment of the OSCE stations was required following application of OWSAT could then be presented to the tool user at the conclusion of the survey, based on the responses provided. Steps involved in the conversion from paper-based to online format of the tool creation process are outlined in table 6-2.

Table 6-2: Steps in conversion of OWSAT to electronic format.

1. Creation of an introduction and welcome script.

2. Inclusion of several baseline demographics questions.

3. Decisions regarding formats within each question e.g. single answer multiple-choice.

4. Provision of free text options including size of comment boxes.

5. Inclusion of skip logic decisions within the survey, e.g. answer no and go to next full question, answer yes and move to deeper questions with free text entry option.

6. Inclusion of any additional questions from results of database testing process.

7. Choice of font for online version.

8. Further opportunity to explore wording of each question, with specific emphasis on any labels included e.g. 'student' or 'candidate' which may be context specific.

9. Inclusion of any graphics and other options e.g. forward and back arrows, duration of survey monitor.

10. Concluding statement following completion of survey.

Moving from a paper-based format to an electronic version in the iterative process of refining the tool was considered necessary due to the need to test a large database of OSCEs for the purposes of identifying any further inclusions or exclusions. This step was undertaken with the expectation that the comments provided in the text boxes could be collated as a testimonial to the experience of using the tool in a local setting and the errors identified, both in the applied OSCE station as well as the tool itself. Use of an electronic version of the tool allowed for downloadable aggregated reports of the types of errors and proposed solutions recorded as the tool was tested against the OSCE station database.

During the conversion to an electronic format changes were made to the tool as required, wherever gaps in the error detection components or issues with clarity or sequencing were identified. These changes related to: 1) additional questions to cover errors that were unlikely to be detected using the options available in the current iteration of the tool, 2) changes in the appearance of the tool for improved readability, 3) additions to the number of answer options, including the addition of options beyond the binary yes/no, and 4) changes to the skip logic, the rules within the software directing to the next question in the sequence to allow for the collection of other information. For example, the addition of a question relating to authenticity was identified through a haematology station involving clinical reasoning and physical examination of a simulated patient. Whilst otherwise sound, the initial description of the patient's presentation may have misled the candidate, both from comprehending the life threatening nature of the disorder, as well as the likelihood that the patient would have been alive long enough to present to the emergency department given the chronology of the symptom development presented in this case depiction. A depiction that is not realistic could easily send a good candidate in the wrong direction when sourcing a diagnosis that begins with key features in the candidates' instructions. Most of the changes made during this phase related to the wording of the tool questions rather than changes to the overall content. Additional changes were related to the use of the Qualtrics® formatting and sequencing. A summary of the key changes is found in Appendix C.

The testing of OWSAT against the first twenty OSCE stations pulled randomly from the sample of hard copy and electronically stored databases led to an immediate awareness of multiple modifications required in both format and content of the online tool version. These repeated changes led to four different versions of the tool used within the first twenty stations tested from the collection. An addition of a question providing opportunity to include recommendations within the tool itself, allowed for ongoing tool testing without the need for constant modifications (question 18). Recognition of the need to update the number of questions listed in the introduction took place after several tool modifications. The information regarding the number of items or questions is to manage the tool user's expectations. At the conclusion of this exercise, there was a total of twenty questions including the final summing up question at the conclusion of the survey. Testing of the tool continued using 100 stations from the available database; however, no new questions were added to the tool beyond twenty questions, as the size of the tool risked becoming too long and less feasible. Ten modifications were made to the original on-line tool, with other questions stemming from the original set of questions, inviting further details to clarify these responses. Beyond those ten modifications, consideration of the use of a qualitative research approach to explore station writers and reviewers opinions on further inclusions to the tool was considered, but was felt to be beyond the scope of the current research project.

There were three key milestones reached in the search through the database of OSCE station for errors to be included in the OWSAT. Firstly, an endpoint was reached when twenty items or domains were reached, which occurred after review of the first seventeen OSCE stations from the database using the OWSAT. During the search for these twenty items, the decision to include each of the items in the twenty questions was considered to be non-controversial. This opinion was reached irrespective of whether the inclusion of a question was derived from the earlier version of the tool or added during the search. Secondly, additional issues were raised by stations reviewed from the database between station 18 and station 67. Whilst no modifications were made to the tool, a list of other potential questions was collated for consideration whether modification or additional information supplied within the current questions would address the concern. Finally, testing using the OWSAT continued until 100 stations had been reviewed. The stations beyond station 67 did not produce any additions to either the questions list for consideration or immediate modifications of the tool. Although errors did exist in these stations, the current tool and further discussion lists held items that already included material to raise awareness of these errors. Testing of the database was considered to have reached the desired saturation point.

A formal statistical validation process such as identifying the number of errors detected pre-and post-use of the OWSAT on the stations in the database was not possible. The purpose of the database testing was to detect errors using my own expertise and experience of station-level problems both vicarious and real. Error detection was dependent on my ability to recognise absent or incorrect information and predict the potential negative impact of these errors in the station scripts. For some stations in the database, this process was assisted by matching flawed stations with examiner and simulated patient feedback provided post-exam and isolating the specific content leading to the negative reports. Collating statistics relating to the number of errors identified with, or without the OWSAT was not the purpose of this exercise. Testing the database for content to be included in the tool for maximum utility, the ostensible saturation point of this exercise was relevant to the tool development. Quantifying error numbers from this database whilst interesting to contemplate was not required for this study and was of dubious predictive value given the sample size. Some stations contained multiple errors within the same domain, for example more than one component of the station information, the candidate's instructions and the simulated patient script containing items that were not authentic, highlighting the need for the narrative options within the tool.

Despite an inability to formally quantify the number of errors using the tool, an opinion on the quality of a station and the degree of modification recommended was anticipated and included in the tool. The final two questions of the OWSAT provided an opportunity to reflect on the overall quality of the station, based on a station writer's or reviewer's previous responses. Consideration of the quantity of errors identified or the negative consequences of particular station errors uncovered using the tool direct the tool user to advise whether the station met one of four criteria for further action. The actions for station writers using the tool are that the station: 1) required minor or 2) major modifications, or, the alternative options of 3) no modifications necessary or 4) retirement. Station retirement was a reasonable option for a reviewed station that had been previously used and was of poor quality, or, required a recommendation to the station writer that this station was unsuitable. These four categories were provided at the conclusion of the tool questions.

The link to the online survey using the free survey creation program Qualtrics® is provided here: https://qasiatrial.asia.qualtrics.com/SE/?SID=SV_6rsh38OssMPDAMt. A pdf version of the initial online tool is found in Appendix D. A pdf version of the final version of the OWSAT followed by the skip logic instructions used in the online version of the survey is found in Appendix E.

The rationale for the inclusion of each question may not be immediately apparent to the tool user, requiring additional communication for clarification. The questions, the rationale for inclusion and any addition communication for clarification provided are located in Appendix F. Further steps regarding improvements to OWSAT both performed and contemplated will be discussed in the following section of this chapter.

6.3 Step 10: Finalisation of tool including suitable graphics and formatting to improve utility.

Utility of the tool required attention to the use of font, the appearance and size of comment boxes, addition of measures to assist the reader to know how far along they were in the duration of the survey and the use of forward and backward buttons to enable changes to be made to answers, to allow for changes of mind and reflection, to reduce user frustration with the tool format. An additional component still under construction is the consideration of how to provide a mechanism for feedback to the OSCE station reviewer. The gathering of information including name and email address at the beginning of the tool provide an opportunity to create a processing option whereby a report of the user's responses is emailed to them at the conclusion of the survey. This aspect of the tool development was untested at the conclusion of the database testing and will be considered under future directions for this project. Some form of feedback mechanism to the reviewer or station writer is essential to reinforce the acknowledgement of station-level errors and permit revision and rehabilitation of a flawed station.

6.4 Chapter conclusion

This chapter outlined the final phase of the design-based research approach to the creation of a tool for aiding OSCE station writers and reviewers to improve the quality of assessment. Experience with use of an online software program known as Qualtrics® supported the choice of this platform for the OWSAT tool. An iterative process was then undertaken to create, improve and test the tool to saturation whereby no further adjustments or additions to the tool were considered. The following chapter will examine the findings, limitations and possible future directions for this research project.

Chapter 7 – Discussion and Conclusions

7.1 Introduction

'The overarching purpose of this work is to challenge those who undertake OSCE research to look beyond traditional psychometric issues.'

(Hodges, 2003a, p. 1135)

In accepting Hodges' challenge (Hodges, 2003a), it is the intent of this research to create an awareness of the many errors which arise during the creation of stations designed to assess aspects of clinical performance. Through this research and presentations at national and international conferences in medical education, and subsequent publications of abstracts in conference proceedings in *Medical Education* journal, a discussion has been initiated about errors in OSCE assessment. A summary of the conference presentations and key messages is provided in Appendix G.

Using a holistic approach to many different aspects of the OSCE station scripts and processes, this project draws on Hodges' three defined discourses in the history of OSCE since its inception in 1975, that of the 'Millers' pyramid and performance' discourse, the 'Cronbach's alpha and psychometric' discourse and the 'Taylorism and production discourse'. Highlighting the possible impact of OSCE station-level flaws on the utility of the OSCE creates an imperative for mandating an effective quality improvement cycle for the process of OSCE station writing.

The OSCE Writers [and Reviewers] Station Analysis Tool (OWSAT) was created to assist reviewers of OSCE stations in a quality improvement assessment framework. This OSCE error-checking tool was designed to support item writers to become aware of poor content or processes that may impact on the performance of a station as a measure of a candidate's clinical competency.

This chapter will provide a review of the work undertaken in this design-based research project. The aims and objectives of this work and the findings linked to the creation of OWSAT will be reported. The relevance of the literature reviewed and how this research contributes to the body of knowledge relating to OSCE assessment will be presented, along with the limitations of this work. Steps undertaken towards validation of OWSAT as a tool will be detailed, and the conclusions of this work, including the significance and future directions, outlined.

7.2 Review of work undertaken

Chapter 1 introduced the topic of clinical competence within the medical profession. Emphasising the importance of using direct observation of clinical skills for assessment purposes, this chapter provided an introduction to the popular simulation-based assessment format, the Objective Structured Clinical Examination (OSCE). The importance of valid and reliable assessment was introduced along with the challenges faced in determining the quality of an assessment. The aim of this research was defined in this chapter as 'to create a tool to aid OSCE station writers and reviewers to identify flaws affecting the ability of the candidate to perform the task'.

The literature review in Chapter 2 explored the vast body of work on the topic of OSCEs. Psychometrics, the predominant language used to describe the quality of assessment and the types of errors encountered in the literature were described. The use of systems to detect errors in the health setting, the concept of assessment literacy and the importance of language in the writing and development of OSCE were discussed through an analysis of the available published research. A gap in the literature relating to explicit acknowledgement of the issue of errors in OSCE station writing was identified. Methods used to address this problem, other than through faculty education, were not evident. The discourse analysis of the history of OSCE, as presented by Professor Brian Hodges, provided insight into changes in research focus over the four decades since the first publication on the OSCE format. Results of the literature review into the benefits and criticisms of the OSCE format, the psychometric principles used to determine the quality of clinical assessment, and a systematic approach to errors were presented in Chapter 2. The significance of a German-language article relating the success of a quality initiative in Heidelberg, including the introduction of a tool for reviewing OSCE stations, faculty development for assessors and feedback to station writers was also discussed in this chapter.

Chapter 3 focused on methodology, and the rationale behind the choice of design-based research to address the problem of how to identify errors in OSCE stations was discussed. The decision to develop a tool using the iterative design-based research approach was introduced in this chapter. Steps undertaken to obtain ethics approval and the three phases of the research project were outlined.

In this research the development of a tool was undertaken using the iterative approach incorporated in design-based research methods in order to revise and improve the comprehensiveness or completeness of the tool content. This method was adopted after consideration of traditional qualitative and quantitative methods, mixed methods and innovative alternatives in the search for a solution to OSCE station-level errors. Design-based methodology is a practical approach addressing the gap between knowledge and implementation. The method involves identifying and clarifying the problem, reviewing the literature and available resources to explore the topic and potential solutions, and the use of an iterative approach to develop and implement a solution. However I am aware of the criticisms of design-based research, which lies in the descriptive studies domain of research. Criticisms include concerns about the validity of this approach, the potential loss of objectivity when the researcher is embedded in the research, and the failure of solutions to translate beyond the local context.

The first phase of the design-based research project was described in Chapter 4. Using an iterative problem-solving approach phase one explored the problem of errors in OSCE, the content of post-OSCE debrief reports and the literature relating to OSCE station-level flaws. The first step was transforming the many different types of errors expressed in the examiner and simulated patient post-examination meetings into a catalogue of potential errors. Phase one concluded with the creation of a list of OSCE questions to consider during station development, covering multiple domains of process and content of station documentation including instructions to candidates, examiners, simulated patients and technical support staff.

Phase two, described in Chapter 5, was devoted to the creation of the OSCE Writers and Reviewers Station Analysis Tool (OWSAT) and peer review of the early iterations of this instrument. The list of questions for consideration when writing or reviewing OSCE stations underwent modification, emerging as the earliest version of the OWSAT. Phase two included an investigation of clarity and formatting options to ensure that the tool captured all the different aspects of the writing process that might contribute to a poor quality station. From the first public presentation of the tool in Quebec City, Canada to the subsequent presentation at a workshop in Ottawa, Canada the following year, the OWSAT underwent many modifications and was subject to peer-review both locally and internationally. The details of the modifications and examples of the various iterations of OWSAT are included in Chapter 5 and relevant appendices.

In Chapter 6 the third and final phase of this iterative process was outlined. Converting a paper-based questionnaire to an electronic format using the online survey program Qualtrics® was the first step undertaken during phase three. The OWSAT was then tested against a database of OSCE stations, sourced from institutions in North America, Europe and Australia for the presence of errors that fell outside the initial set of questions and risked non-detection by the tool user. Additional questions determined from this testing process were included in the OWSAT. This process continued until a saturation point, where no further alterations to the tool content were identified from the database testing process. The modification of the wording of some questions and the insertion of additional content uncovered during the testing phase were the final steps in this iterative process.

Discussion

7.3 Research Aims and Questions

Rather than a single research question this project had an underlying aim and a series of questions were generated in relation to that aim. These questions included the following:

1. *What aspects of the OSCE item writing process are prone to errors that undermine the quality of this assessment format?*

2. *What elements of known best practice in OSCE station writing should be included in a tool to aid OSCE writers and reviewers to improve the performance of the OSCE?*

3. *What steps are necessary to create a useful tool to assist OSCE writers and reviewers to identify potential errors in the writing phase of the assessment process?*

4. *Does the tool make the invisible visible with respect to flaws within an OSCE station enabling writers and reviewers to identify errors during the pre-exam quality improvement processes?*

5. *What is the utility of a tool to enhance the quality of OSCE stations in assessment?*

These questions will now be discussed.

7.4 Findings with respect to research aims and questions

1. *What aspects of the OSCE item writing process are prone to errors that undermine the quality of this assessment format?*

It is clear that all aspects of the OSCE writing process are prone to errors that can undermine the quality of the assessment. Each station has a task list for the candidate, examiner, simulated patient and technical support or set up team, all of which may contain errors. Errors preventing the candidate from understanding or implementing the required task, despite possessing the ability to perform competently will lead to decreased reliability and validity of the assessment. Best practice involves respect for protocols relating to patient safety and use of evidence-based medicine for investigation and management decisions in medicine. Standards relating to patient safety include adherence to hand hygiene standards. The opportunity for the candidate to wash his or her hands or use alcohol-based hand rub before or after each patient contact is essential. Failure to provide these at a time of assessment implies a lack of prioritisation of patient safety in the minds of candidates and others involved in the assessment, including examiners and simulated patients. Regardless of any emphasis during the education program, messages sent through assessment station design will have a potential negative impact on the future actions of candidates (Hays, 2008).

2. *What elements of known best practice in OSCE station writing should be included in a tool to aid OSCE writers and reviewers to improve the performance of the OSCE?*

This question was addressed in the creation of the OWSAT and the questions included in the final version of the tool design. Elements included in the tool related to the choice of OSCE for assessment, communication of task and timing and ensuring that sequencing and other station details aid authenticity. In addition, the tool considered the safety of simulated patient and the candidate in the examination as well as the internal consistency of the station details. Whilst the marking tool format, whether a checklist or global rating scale format was not emphasised, the requirement for the marking sheet to match the station objectives, and be relevant to the task details on the candidates' instruction sheet was considered within the tool. Blueprinting the station against the candidate's curriculum to ensure relevance of the learning objectives to the year level was also included in OWSAT.

3. *What steps are necessary to create a useful tool to assist OSCE writers and reviewers to identify potential errors in the writing phase of the assessment process?*

Chapters 4, 5 and 6 contained details of the three phases and nine steps of the design process leading to the current version of the OWSAT. Identification of errors for inclusion in the tool was initially based on observed and reported existing errors within post-OSCE debrief meetings with some additional content based on the literature review detailed in Chapter 2. Further content was contributed following two peer-reviewed meetings and the testing of the tool against the database of OSCE stations. Consideration of the platform, formatting of questions and options for responses and length of time taken to use the tool were additional steps involved in the tool-creation process.

4. Does the tool make the invisible visible with respect to flaws within an OSCE station, enabling writers and reviewers to identify errors during the pre-exam quality improvement processes?

Attempts to determine the answer to this question relating to the efficacy of the tool were undertaken at two focus group workshops held in Melbourne and Ottawa at key health professional education conferences. There may be possible differences in efficacy between those with extensive experience in OSCE writing and development and novices with low assessment literacy. Further research is necessary to answer the question of whether the tool works to uncover flaws in OSCE stations. Determining the context where the efficacy of the tool is acknowledged and maximised will be a key question for subsequent research into quality improvement of OSCE stations using the OWSAT.

5. What is the utility of a tool to enhance the quality of OSCE stations in assessment?

The utility of the tool in different contexts and by academics with variations in assessment literacy or experience in OSCE writing will require further investigation and is beyond the scope of this project. During this project the OWSAT was demonstrated and tested in workshops attended by an international audience. The positive reception from participants is encouraging; however, a formal approach is necessary to answer this question.

7.5 Research Conclusions

This research is fundamentally concerned with the quality of assessments of clinical competence or performance in a simulated patient encounter using the OSCE format. There is a paucity of articles in the literature relating to the inclusion of discussions about errors or processes that undermine the quality of assessment. The reticence to publicly disclose flawed OSCE station details has led to the perception that errors are rare or insignificant in their effect on assessment validity. It is possible that concerns regarding litigation from failed students may have facilitated silence regarding station-level flaws.

As discussed in Chapter 2, a normalisation of deviance exists in medical assessment where we accept examinations with poor reliability. Tolerance of poorly performing OSCE stations aligns with an evaluation that our assessments of clinical competence are good enough even if the Cronbach's alpha measurement lies somewhat below 0.8. Schuwirth and van der Vleuten are critical of 'a scientific model explaining so little of the total observed variance but which is used to predict such important future performance' (2006, p. 297). If multiple stations contain significant flaws within a single examination, the measurement of internal consistency may in fact be high, with the quality of the individual components consistently low.

There are remarkably few descriptions of OSCE stations that are withdrawn due to concerns about the fairness or validity of decisions based on the results of the observed encounter in that station. Of note, only Auewarakul and colleagues (2005) from Thailand presented a comprehensive account of a series of clinical examinations where multiple stations per cohort failed to meet their determination of an effective station. Parallels with the patient safety literature and the effect of the culture of silence on a failure to propagate vital information to prevent repeated deaths from the same error are evident (Donaldson et al., 2000; Taylor-Adams et al., 2008). This knowledge of failure prevention translated from the aviation field into medicine provides an exemplar for benefiting from a sharing of adverse events in assessment, including performance-based formats (Parry et al., 2012). The first step in this process is acknowledgement and sharing of the errors that are identified in both station-level writing and OSCE processes to reduce the necessity for all institutions utilising the OSCE format to learn only from their own experience.

Design-based research is used to answer questions relating to why a problem exists, what is needed to make an effective solution and which aspects of the process of designing and implementing a solution can contribute meaningful information for future design-based researchers (Educause, 2012). Kelly (2004 p. 116) states that design-based research should 'produce an artefact that outlasts the study'. It is clear that within Kelly's (2004) broad description of what constitutes an artefact, the tool would meet the criteria for a product of design-based research. A further recommendation for output from design-based research is that the artefact gets used, and modified by future users beyond the initial research project. Further research into the utility of the tool, how it is used and what modifications are done to improve it is the next stage of the development of the OWSAT.

7.6 Limitations and Suggestions for Future Research

The aim of this research was to improve the assessment of clinical competence using the OSCE format, by aiding station developers and reviewers to identify station-level errors. Pursuing a solution for improving the quality of OSCE station writing resulted in development of a tool aimed at assisting OSCE station writers and reviewers to be more aware of these errors.

This project led to exploration of best practice assessment principles as they apply to the assessment of clinical competence in the format known as OSCE, in particular, the inclusion of elements within the item writing process that are flawed, providing the potential to undermine reliability, validity and impact on learning of the candidate. This project focused only on the item writing process for OSCEs, whilst recognising that the OSCE is only one measure for assessing competence in clinical skills. An assumption was made that the literature from both the medical education field, as well as the education field, generally contained sufficient information to assist in the collation of a list of elements that must be considered in the item writing process to ensure that assessment errors based on station design are minimized. The contribution of other processes and human factors to the reality of flawed OSCEs and failed OSCE stations, such as examiner bias, is acknowledged but does not form part of this study.

A major limitation of this project is that only one person undertook the literature review to assemble elements for inclusion in the tool and contributed to the search for errors within the OSCE database. Similarly, within the workplace-based assessment literature, the effect of a single rater is a useful contribution to the gathering of evidence in support of a pass or fail judgement regarding a student's performance, but this is insufficient to mount a convincing argument in favour of validity (Yeates et al., 2013). The language, processes and validity expectations from a clinical examination are highly culturally dependent. A literal interpretation of station information has the potential to create an error where perhaps one should not exist, given a shared language to describe a task, for candidate, examiner or simulated patient. The perception created by a single individual with a particular cultural background may differ significantly from that observed by another person, resulting in fewer or more errors being identified when reviewing OSCE station information.

The testing phase of the tool was dependent on my ability to identify flaws in the OSCE stations and interpreting them within the structure of existing domains or questions looking for outliers. Researching the topic of OSCE errors for this project has substantially increased my assessment literacy. An OSCE writer or reviewer with more limited assessment literacy may still fail to identify existing errors, even with the assistance of the OWSAT, as they are unable to foresee the educational impact of a particular wording or missing additional task within a station. The tool may require additional explanatory information for users with low assessment literacy or limited experience with the language of the OSCE. A pop-up glossary tool may be useful in an electronic version of the tool.

The current version of the tool does not provide personalised feedback to an OSCE reviewer other than their own interpretation following the completion of all the available questions within the online version of the OWSAT. Ideally, feedback provided to the tool user based on responses to particular questions could specify whether identified errors would undermine the validity, reliability, feasibility or educational impact (see Table 5-2, p. 73). The ability to provide an emailed response direct to the reviewer or OSCE author to facilitate structured feedback and corrections to the flawed station would improve the utility of the tool.

Evidence of the capacity of the tool to improve the quality of the OSCE cannot be verified within the current scope of this research project. Gathering psychometric data to determine reliability using generalizability theory or Cronbach's alpha for OSCEs at a number of sites pre- and post- introduction of the tool would provide a useful piece of information to assist with determining validity. The use of a qualitative research approach to assess opinions from OSCE station writers and reviewers regarding utility and acceptability of the tool in the quality improvement process and gathering pre- and post-OWSAT opinions of examiners and simulated patients regarding the quality of OSCE stations in an examination would also be a useful study for determining the value of this intervention. Finally gathering student opinions regarding OSCE quality at centres where the majority of stations are written and reviewed by the same team of people across different year levels or in sites where students undertake more than one OSCE in a year, or where there is a significant number of repeating candidates e.g. vocational clinical examinations would be an additional element of validity evidence required for determining the tool's impact. Controlling other variables which can improve the overall examination reliability is challenging, given the effect of candidate related variables, examiner and venue factors, as well as other unknown contributions to the variability of station and examination results. The Heidelberg study published the results of three interventions, the introduction of their OSCE reviewing tool, examiner training and feedback to OSCE station writers, attributing improvements in the assessment program to the combination of these changes (Schultz et al., 2008). They did not attempt to attribute any specific measure of contribution from the tool alone to their improved OSCE performance.

Gathering evidence for the purposes of validation of OWSAT is a clear direction for future research into OSCE errors. No defined threshold for validity evidence exists. Convincing others that the tool works for the stated purpose requires sufficient collating of observations, data and narrative for this purpose. A focus of future tool testing will be to investigate whether different users attain the same results applying the tool on the same set of stations, the response aspect of validity. Identifying the effect on other variables, such as the psychometric analysis of OSCE performance pre- and post- OWSAT introduction, is a separate end point from the numbers of errors detected using the tool and may demonstrate a longitudinal benefit of the tool in quality improvement of the assessment.

7.7 Critique of thesis

Kaufman and Keller (1994) speak of the threat and promise of evaluation. The promise is the potential for information obtained to be used for a continuous quality improvement cycle, whilst the threat is that ‘performance data will be used for blaming and not for fixing and improving’ (Kaufman & Keller, 1994, p. 371). The purpose of the tool is to allow station writers to ask a series of questions about their creative output, with the intention of identifying aspects that need further development or refinement. There is the potential for the tool to be misused to embarrass station writers. It should instead form part of a quality improvement program, using a systems-based approach to identify flaws and remedy them if possible.

The identification of errors within a station does not automatically lead to correction of these issues. The evidence from the post-OSCE debrief meeting notes indicated a lack of institutional capacity or motivation to undertake meaningful changes to the inclusion of known problem stations in future examinations. If effective the tool will support motivated individuals to explore quality improvement in station wording and may empower reviewers to speak about observed errors using the formalised structure of the tool. This socio-material exploration of the interaction of the tool in context cannot fully be realised without further study of the effect and effectiveness of the OWSAT.

As discussed in Chapter 3, full validation of this tool did not fall into the scope of this project, but does map the pathway for the future (Ilic et al., 2014; Schou et al., 2012). Using Beckman's (2004, p. 973) interpretation of the Standards for Educational and Psychological Testing the criteria of validation '1) content, 2) responses, 3) internal structures, 4) relationship to other variables, and 5) consequences' provides a guide to areas where information relating to the development and utilization of the OWSAT may contribute to understanding the validity of this tool. An alternative validation pathway is provided through Guba and Lincoln's criteria of credibility, transferability, dependability and confirmability (Guba & Lincoln, 1988, p. 111). I believe that the perceived acceptance of the tool when demonstrated at medical education conferences provides some evidence towards the credibility criteria but this aspect should be formally explored in future studies. Through the process of reviewing the literature, identifying content from the transcripts of post-OSCE debriefs and adding further content from the testing of a large number of stations for the presence of additional content for the OWSAT, the criteria of content for validation using the Standards of Educational and Psychological Testing, is also addressed. The other criteria are beyond the scope of this project; however, validation of OWSAT remains a longer-term goal should the tool meet other criteria such as utility and efficacy.

7.8 Conclusion

In the realm of medical assessment, the role of the OSCE as an established method of assessing clinical competence outside the workplace is undisputed. First described by Harden in 1975, the use of the OSCE has rapidly spread across the globe adding considerably to the reliability and validity of assessment programs and providing a steady source of topic for research and publication. Despite widespread acceptance the OSCE has met with criticism in the literature for issues including authenticity and validity, particularly when undermined by station level errors.

The aim of this research was to contribute to improving the quality of the assessment of competence using the OSCE format, by aiding station developers and reviewers to identify station-level errors. The guiding research question was:

> *What aspects of the OSCE item writing process are prone to errors that undermine the quality of this assessment format and how can these be overcome?*

However, given the iterative nature of design-based research, this question evolved along the journey to include the following:

1. *What elements of known best practice in OSCE station writing should be included in a tool to aid OSCE writers and reviewers to improve the performance of the OSCE?*

2. *What steps are necessary to create a useful tool to assist OSCE writers and reviewers to identify potential errors in the writing phase of the assessment process?*

3. *Does the tool make the invisible visible with respect to flaws within an OSCE station, enabling writers and reviewers to identify errors during the pre-exam quality improvement processes?*

4. *What is the utility of a tool to enhance the quality of OSCE stations in assessment?*

Using a design-based research approach, a tool to aid in the quality improvement of OSCE station writing has been developed. Using a three phase iterative process, this thesis has outlined the steps and decisions through the development of the OWSAT – the OSCE Writers Assessment Tool. This process has included much iteration, peer review at national and international conferences, and application of the tool against a database of OSCE stations. Flawed OSCE stations were not unique to one institution, nor were they confined geographically or within levels of medical education. Errors were detected in many of the stations in the OSCE database during the systematic analysis of the key components of the station details using the OWSAT. This testing phase worked towards saturation point where no more new error types were detected.

This project has met the aims by creating a tool to aid OSCE writers and reviewers to identify and correct errors affecting the validity, reliability, feasibility and educational impact of the assessment of clinical competence. Whilst not yet validated, the OWSAT contains questions highlighting aspects of OSCE writing that require reflection by the OSCE station writer or reviewer in search of improved station performance and overall assessment quality. It is anticipated that OWSAT will assist OSCE station writers and reviewers to identify errors at the writing or reviewing stage of the assessment process, and consequently enhance the systems that use OSCE assessment approach to determine competence in the medical profession. Peer review provided an overall positive reception for the drafts of the tool. It is anticipated that further refinement and dissemination of the tool will enable global acceptability and utility for improving quality in the OSCE format for the assessment of clinical skills at both the undergraduate and postgraduate levels. An opportunity to test the validity of the OWSAT is highly anticipated by this researcher.

Glossary of Abbreviations

SP = simulated patient

GP = general practitioner

GI = gastrointestinal

PR = per rectum

ID = identification

CXR = chest x-ray

CPR = cardiopulmonary resuscitation

BLS = basic life support

ECG = electrocardiogram

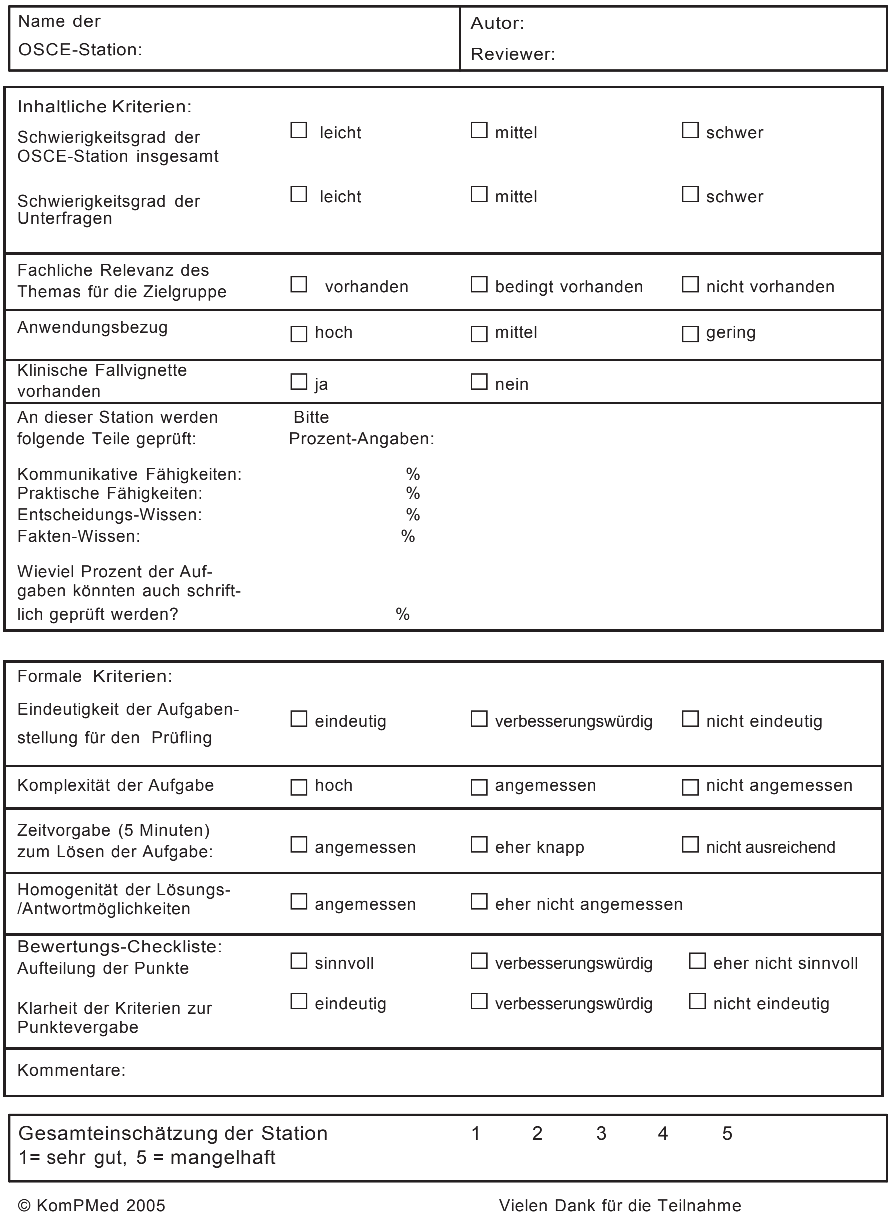
SVT = supraventricular tachycardia

NHL = non-Hodgkin’s lymphoma

Min = minute

Appendices

Appendix A: Heidelberg OSCE Station Review Checklist



Name der OSCE-Station:	Autor: Reviewer:

Inhaltliche Kriterien:			
Schwierigkeitsgrad der OSCE-Station insgesamt	□ leicht	□ mittel	□ schwer
Schwierigkeitsgrad der Unterfragen	□ leicht	□ mittel	□ schwer
Fachliche Relevanz des Themas für die Zielgruppe	□ vorhanden	□ bedingt vorhanden	□ nicht vorhanden
Anwendungsbezug	□ hoch	□ mittel	□ gering
Klinische Fallvignette vorhanden	□ ja	□ nein	
An dieser Station werden folgende Teile geprüft:	Bitte Prozent-Angaben:		
Kommunikative Fähigkeiten:	%		
Praktische Fähigkeiten:	%		
Entscheidungs-Wissen:	%		
Fakten-Wissen:	%		
Wieviel Prozent der Aufgaben könnten auch schriftlich geprüft werden?	%		

Formale Kriterien:			
Eindeutigkeit der Aufgabenstellung für den Prüfling	□ eindeutig	□ verbesserungswürdig	□ nicht eindeutig
Komplexität der Aufgabe	□ hoch	□ angemessen	□ nicht angemessen
Zeitvorgabe (5 Minuten) zum Lösen der Aufgabe:	□ angemessen	□ eher knapp	□ nicht ausreichend
Homogenität der Lösungs-/Antwortmöglichkeiten	□ angemessen	□ eher nicht angemessen	
Bewertungs-Checkliste: Aufteilung der Punkte	□ sinnvoll	□ verbesserungswürdig	□ eher nicht sinnvoll
Klarheit der Kriterien zur Punktevergabe	□ eindeutig	□ verbesserungswürdig	□ nicht eindeutig
Kommentare:			

Gesamteinschätzung der Station 1= sehr gut, 5 = mangelhaft	1	2	3	4	5

Vielen Dank für die Teilnahme

Abb. 1. Checkliste für den Prae- und Post-Review von OSCE-Stationen.

(Schultz et al., 2008, p. 671)

English translation of Heidelberg Checklist (translated by K. Brotchie)

Name of OSCE-Station:	Author: Reviewer:

Content Criteria:

Difficulty of OSCE Total Station ☐ easy ☐ medium ☐ difficult

Difficulty of sub-tasks ☐ easy ☐ medium ☐ difficult

Choice of topic for this cohort ☐relevant ☐partially relevant ☐not relevant

Practical relevance ☐ high ☐ medium ☐ low

Clinical Case Vignette available ☐ yes ☐ no

At this station the following are assessed: (Please provide percentages)

Communication skills: %

Practical skills: %

Clinical reasoning: %

Factual knowledge %

What percentage of tasks could also be tested in a written? %

Formal criteria:

Clarity of the task ☐ clear ☐ needs improvement ☐ unclear

Complexity of the task ☐ high ☐ appropriate ☐ highly inappropriate

Timing (5 minutes) for completing task: ☐ fair ☐ too short ☐ insufficient

Case diagnosis ☐ suitable ☐ unsuitable

Checklist evaluation:

Choice of marking criteria ☐ helpful ☐ needs improvement ☐ unhelpful

Clarity of marking criteria ☐ clear ☐ needs improvement ☐ unclear

Comments:

Overall assessment of the station: 1 2 3 4 5

1 = very good, 5 = poor

Fig.1. Checklist for the pre-and post-review of OSCE stations (Schultz et al., 2008, p. 671)

Appendix B: Early Clinical Years' OSCE debrief notes. November 2011

Summary of Transcript of meeting post-OSCE (De-identified) - Audiotaped and transcribed by Dr Kathy Brotchie.

Request for debrief – work through case by case to hear what you thought about it. Be succinct. We want both SPs and examiners to comment. Comments to be about the case, for example - was it expressed logically, were there any hiccups etc…

Station 1. Diverticulitis – person with left sided belly pain. Generally done well, but should be. This diagnosis shouldn't be too difficult. Some history taking was not done well. Surprising given they did history last year. Many students when talking to the patients stated that diverticulitis was a normal process, not a diseased one. Students got hung up on IV drug use and alcohol and not on bowel habit to establish a change from normal. Some spent too much time asking permission. E.g. Now, I am going to ask a few questions on family history – is that OK. Dietary history went missing. There was not enough of that.

Station 2. Single straightforward station. History. Great instructions, both examiner and student instructions were good, including SP instructions which were also good. Well done. Students had just done a GI history previously; we were surprised that they didn't just slip into it. This was a good station to discriminate – there were a few critical errors made – including flaws in history taking – we haven't got it through to them that if they ask an open ended question e.g. Have you had any bleeding, then follow up immediately with a closed question e.g. Have you had any nose-bleeds you get a 'no' as only the closed question will get answered. This will make students miss a history of ?PR bleeding.

Station 3. This station is very condensed. We had to check the ID of the student, which was on them and had to get too close. There were another two interruptions – which lost 45 seconds of the station time just with the introduction. At five minutes we have to ask about diagnosis, the student then had to read results and think about them. You don't need to have an introduction by the examiner in the station –the information is already in the stem and takes up valuable time from the students' task. Far too many points in the marking sheet were for establishing who they are and introducing themselves, more points needed for history of presenting complaint (HOPC). The marking sheet should have more options. There were also issues with this station in terms of structure. The odd candidate brought up – 'have you been to the doctor', which then gets out of sequence, the SP could either lie or say they hadn't but then later you say you have. The station forces the SP into not telling the truth. Station then produces results so clearly they have seen the doctor. The examiner then has to give the relevant findings and there were issues to get back on track. Many students made the diagnosis with only half the information. One of the questions from the SP was unclear – 'When I go home, what can be done to stop this happening again?' – not sure whether what was wanted was what can we do as doctors, or what can the patient do to help themselves. Station needed rewording.

Station 4. Was it deliberate that there wasn't any hand washing in the stations? Interesting that we had a discussion about that yesterday. Dr X decided to only have hand washing in non-history stations. (Problem is for SPs in history stations after examination ones – don't know if the student did actually wash their hands in the examination station).

This station is a cranial nerves examination. Issues - Sim patients need to be very consistent. If students are not careful enough in their technique they will miss the ® sided hemianopsia. The afternoon groups were not considered to be as good as the a.m. group. Often missed the ® sided hemianopsia. One student declared patient was blind in ® eye despite reading the Snellen chart with both eyes. Patient had a? Bitemporal hemianopsia.

There was too much history in this station. Patient states – 'I have a problem with my eyesight; I nearly had an accident last week'. In one of the circuits – we didn't give the history unless the student asked and then they were given it.

Cranial nerves II, III, IV, VI were to be examined. Most students wanted to be prompted to give the findings – the examiner felt they had to prompt them despite the wording of the station.

Station 5. Respiratory examination. This is a fairly busy station – students had to do the examination, give findings, and read the x-ray. Students became very flustered. Too much time was spent on consent- saying what they were going to do. No marks were given for hand washing where some didn't wash their hands and you couldn't mark them down.

Some students didn't do the respiratory rate, or pulse or only did pulse. Not best quality CXR – very difficult to interpret. Recommend using a laptop computer screen to show the image in future.

What are we asking them to do? Summarise normal examination findings? What else can I do to finish – 1 min for CXR interpretation is not the best? I can't see anything. Only 4 or 5 got correct or incorrect findings.

Station 6. Shoulder examination – SP was excellent. Students were convinced he had a painful shoulder. Station was too short – almost everyone finished at 6 minutes. Should have had X-ray in this station not the previous one.

Station was not a very good discriminator. Most didn't waste time with consent. Part of the marks was to explain. The form was good – well set out.

Disappointed with the students' palpation skills – they had trouble finding and identifying the landmarks – many kept hand on the sore spot while they were examining.

Hand wash – should we test or not? Impingement testing was not done often. It is not in Talley & O'Connor, however many got the empty beer can test done. Station was not specific enough in directions. Should they interrupt to get feedback? Not clear to the examiner. Some went onto examining neurology before feeding back.

(Comment from Dr X – Last 2 stations were written last year. Aiming for 1/3 new stations every year.)

Examiners felt that it wasn't really an appropriate station for Year 'X'. Wanted to know when they would realistically examine a shoulder on the wards.

Station 7. Unclear if it was discriminating. There was confusion between basic life support vs. hospital life support using bag and mask. 5 fails, borderline 4-5 by one examiner. Most do not understand about airway. 1 student asked the examiner if the patient (lying on an examination couch) was on the floor – despite the stem stating they were taking history from patient on medical ward when they collapsed. Lots called the ambulance instead of met call or calling arrest team – not aware of hospital location given in stem.

Most finished early.

Need a rest station after a CPR station. One examiner had them do more compressions to fill up the time. Many students were very upset about doing it all on their own. Some didn't do the bag mask at all.
One examiner stated that the guidelines say you are not to spend too much time doing bag mask, just do compressions (not correct – confused examiner thinking of changes to BLS – outside hospital guidelines).

Many didn't check the airway at all. May be just an airway obstruction. Issue of experience and discernment.

This year had introduction of critical error. ? Airway management is critical. Many students failed the station.

At one hospital an anaesthetist has started training all staff and found most are very poor at airway management.

Station 8. Arrhythmia. Obvious diagnosis. Most got it. ECG interpretation was done very poorly. A number of students actually picked limb reversal. The ECG provided was too small. One ECG per time should be on a page. It should be a proper A4 12 lead ECG document, as they would see in the hospitals these days.

The patient was a 25 y.o. with palpitations but the history that they used was what you would use for an 85 y.o. The students should focus on age related questions.

The ECG was too fast. 150 per min. There could be p waves present but this was unclear. The marking sheet talked about ?no p waves.

Examiner noted that if you missed a dot on the iPads and hit the bottom send by mistake it wiped about all the bubbles and you had to remember what the student had already done.

Many students had no concept of what sinus rhythm was. SVT ECG needs a rate of 170 per min with no p waves. Most guessed about limb lead reversal. If they picked that the chest leads were reversed not limb leads reversed they still got masks as the marking sheet talked only about lead reversal.

There was also a box for chest pain and a box for chest heaviness on the marking sheet. Redundant.

Many asked relevant questions but they couldn't get positive points as tick box didn't have them.

Another examiner reported there were p waves in every lead. Delta wave in one. Should have a normal speed on it. Need a big ECG.

The students became fascinated with the variation in heart rates of the two ECGs printed on the one sheet. ?technical problem.

Many students seemed to have a problem with personal space – leaning too forward, too close (? because ECG was too small). Request was for students to be provided a pointer as the students wanted to use their biros to point and would leave marks on the ECG.

Station 9. Pathology – Hodgkin's or NHL lymphoma. This station was a counselling session about the findings of a diagnosis about this condition. The students were given histology report. One examiner pointed to the bottom to give help – just look at the diagnosis. Students wanted to read the whole report but time didn't really permit this. The students assumed that basic bloods had already been done if the patient made it to a lymph node biopsy. Bone marrow should be on the list of choices of further investigations, but CT, LFTs etc were there? One examiner prompted students for basic tests despite it not being in the station examiner instructions. Didn't seem to comprehend this as an issue for consistency.

'Patient's diagnosis is …. ' should be on the stem. The fake scar was a good visual cue. Many didn't seem to know what a normal lymph node should look like. Only 2 or 3 addressed the patient by their name. They are then asked what further tests to request and their reasons for it.

Did hear students weren't able to read the pathology report all the way through.

Dr X – we learnt rapidly through the day there were technical difficulties in this station. Some words in the report or questions were unclear. Some students struggled with the physical architecture of lymph cells. One examiner thought we should change the words to help the international students - did he do this? Also thought path report should be on the door not in the room. Monash instructions not clear. Disaster of a station – issue also with pathology teaching?

Station 10. 75 y.o. with rhythm disturbance on warfarin. All felt it was appropriate to have patient's name on the stem. Student should be able to know the patient's name. Confusion about whether they had to talk about Atrial fibrillation as well as anticoagulation. Poor discriminator. 2 borderlines, all the rest passed. They handled it very well.

Options on the marking sheet were 'not done', 'partially done', and 'covered well'. There should be more options. Some did well but did say some things wrong – which made them difficult to mark. The scoring pad didn't correlate well with student's instructions. Scores relied on patient's prompts and there were conflicts between the SP instructions, scoring sheet and student instructions.

There was nothing in either instruction sheets (student or marking sheet?) (Student or SP instructions?) to talk about being on warfarin for life or the use of alternatives. Better students were disadvantaged by knowing too much. This patient was a well person with atrial fibrillation and no other risk factors and wouldn't have qualified for warfarin under best practice guidelines. SP had to give set questions about the use of warfarin. There were marks for eliciting patient's concerns but the patient was expected to give this through their questions so we were unable to fail any of them. This station needs to be retired.
The students were given an article about atrial fibrillation causing stroke – given a mark for picking that.

Students need education about counselling and personal space. We need to put Mr or Mrs on the door.

Students were meant to be a GP for this station. They had no authority to provide this information to the patient otherwise. Many students introduced themselves as Yr 'x' students rather than get in role and pretend to be GPs.

Appendix C: Changes to OWSAT based on database testing in Phase Three

Station number	Year Level	Task for station	Tool requires modification? Comments	What change did I make to the Tool from this?
1	UEC	Complete history in 8 minutes	Font Comic Sans Serif. Doesn't read well on iPads. Error with display logic for question 15 misdirecting from wrong answer.	Changed font to MS Sans Serif Corrected display logic for Q 15.
2	UEC	Abbreviated CVS exam missing carotids, peripheral pulses.	Error with display logic for Final question - option refinement. Following screen was describing retirement option not thank you screen. Error with wording of Q4. Have you selected the most appropriate assessment format.. Changed to ....Has the appropriate assessment format been selected for all tasks in this station?	Q4. Changed to ....Has the appropriate assessment format been selected for all tasks in this station?
3	UEL	Haematology Examination, clinical reasoning, management,	Issues with authenticity, unrealistic presentation for patient with this particular condition.	Added Question 17. Question regarding authenticity.
4	UPC	Musculoskeletal examination of Knee	Modification to Q6 on resources required. Needs unknown, or information not supplied. Some stations don't include resources information. Tool requires an option for 'marking guide needs discussion'.	Modification to Q6. Needs _not known, or information not supplied. Some assessment writers don't include this information? Unclear why. Also Tool requires an option for 'marking guide needs discussion'.
8	UEC	Paediatrics history Gastro case developmental milestones and interview of mother	Yes. Intro states presence of 16 questions. There are now 17 Q.	Updated Intro to state 17 questions.

Station number	Year Level	Task for station	Tool requires modification? Comments	What change did I make to the Tool from this?
12	PVT	Read hospital discharge letter, request investigation results, define and manage significant elements	Yes. Use of term student is not appropriate for vocational examinations. Recommend use of word candidate instead throughout the tool.	Addition to introduction discussing use of student or candidate to assist with utility of tool in different contexts.
13	UEL	Failure to thrive	No, but discovery of need to fully define acceptable terms within candidate instructions became clear. E.g. use of words to provide advice about likely cause and possible interventions.	Added to thesis additional material regarding complexity of not having a truly shared language for describing tasks in clinical skills.
14	UPC	Hip examination	Yes. Intro states presence of 17questions. There are now 18 Q. plus additional summing up selection options.	Modified introduction to state approximately 20 questions depending on answer choices to assist with tool user expectations.
17	UPC	Blood glucose testing and urinalysis	No, but consideration of spare equipment for station prompts consideration of a tool for identifying OSCE resource requirements as a separate study.	
18	UPC	Upper limb neurological examination	No, but reinforces need for patient safety question.	Considered addition of preamble including sharps containers, and not using neurotips in examination setting, paperclip option preferred to examiner stating sensory examination not required. Complex discussion perhaps for future feedback from tool users.

Station number	Year Level	Task for station	Tool requires modification? Comments	What change did I make to the Tool from this?
33	UPC	Musculoskeletal examination of back	Scope of practice question needed. E.g. Are students required to demonstrate skills in a manner where they are role-playing a different scope of practice than their current status? Issues for identity formation and authenticity.	More discussion with team of tool users required regarding scope of practice and whether this is covered by the authenticity or harm to student questions.
45	UPC	Focused history of presenting problem	Yes	Consideration of terminology used for the candidates task. Does it require education specifically for this terminology?
47	UPC	Communication HIV result	Yes. Sensitive station raising issues of possible wording in a similar station that might cause offense. Political concerns re diversity and stereotyping.	Need to add question regarding 'have you considered gender, diversity and cultural sensitivities in the writing of this station' - also raised at workshop in Ottawa.
48	UPC	Prostate examination	Yes.	Needs a question regarding length of time taken to read the candidate's instructions.
49	UPC	Basic life support	Yes. Need to ensure the equipment used is appropriate for the setting.	May already be incorporated, but could be done better?
51	UPC	Pain history	Yes.	Have you included sufficient details in the candidate's instructions?

Appendix D: Qualtrics Survey – Phase Three, first iteration of online tool

9/12/2014 .... Qualtrics Survey Software

OSCE Station Name ....

Is the author of this OSCE Station known?

○ Yes ○ No

If yes, please enter the name of the Author for this station. ....

Please describe the task or tasks to be performed in this station ....

OSCE Station Evaluation Checklist

Have you selected the appropriate assessment format for each task in this station?

○ Yes ○ No

If No: Describe the task and recommendation for alternative form of assessment

....

Are there adequate resources, equivalent to what would be found in normal clinical practice?

○ Yes ○ No

What additional resources are required for this station?

....

Has the station been designed to allow sufficient time for each task, and for all tasks to be completed.

○ Yes ○ No

Please select the most appropriate response

○ Insufficient time has been allocated for all the tasks to be completed in this station

○ Station requires deviation from best practice to complete task in time allowed e.g. student must take short cuts or other modifications to their usual approach to task. (Please describe)

○ Station has adequate time for all tasks to be completed, however, timing for individual tasks needs modification. (Please describe)

Student instructions indicate how much time is permitted for each task

○ Yes ○ No

Does the order of tasks follow a logical sequence for this station?

○ Yes ○ No

Does this station follow current clinical guidelines/ best practice for the task(s)

○ Yes ○ No

Do the student/candidate instructions provide clear information about the key tasks on which they will be assessed?

○ Yes ○ No

Is there alignment between the task marks and the station objectives?

○ Yes ○ No

Have any potential risks for student, patient or examiner been identified and managed?

○ Yes ○ No

Are the students being assessed on material they would have been expected to encounter in the current curriculum?

○ Yes ○ No

Is there anything in the format or expectations of this station which may have a negative impact on the students' clinical skills practice in the future?

○ Yes ○ Maybe ○ No

Based on the above evaluation this OSCE station requires:

○ Redevelopment ○ Refinement ○ Retirement

https://asia.qualtrics.com/ControlPanel/Ajax.php?action=GetSurveyPrintPreview&T=

Appendix E: Qualtrics Survey – Phase Three, second Iteration of online tool – no feedback available to reviewers

OWSAT Tool - A guide to assist station writers and reviewers

Welcome to OWSAT, the OSCE Writers and Reviewers Station Analysis Tool. This tool has been designed to aid in improving the quality of your OSCE stations. By reflecting on different aspects of the OSCE station wording we hope to assist with identification of flaws that may affect the performance of the station. The tool has approximately 20 questions (depending on your answers) and should take less than five minutes to complete. This time will depend on just how many errors you uncover, and how much feedback you wish to provide regarding station improvements. We respect your choice to use the word student or candidate in your OSCE station to reflect the culture of your assessment.

Please provide the following information. In doing so, we can send you a copy of your responses to the question for your reference.

Name ....

Email..

1. OSCE Station Name

....

2. What is the source of this Station? (Please tick all that apply)

□ Preclinical undergraduate years

□ Early clinical undergraduate years

□ Late or exit clinical undergraduate years

□ Vocation training years

□ Other ....

3. Feedback to the station writer is a key element of the quality improvement cycle. Is the author of this OSCE Station known?

○ Unknown ○ Known (go to 3a)

3a. If known, please enter the name of the Author for this station.

....

4. Please describe the task or tasks to be performed in this station.

....

....

5. OSCE is an expensive format for assessing pure knowledge that could easily be tested using a written examination. Has the appropriate assessment format been selected for each task in this station?

○ Yes ○ No (go to 5A)

5a. if no, describe the task and recommendation for alternative form of assessment.

....

6. The correct tools to perform a task and sufficient consumable items to prevent interruptions and delays on OSCE day are important considerations for setting up this station. Are there adequate resources, equivalent to what would be found in normal clinical practice?

○ Yes ○ No (go to 6a) ○ Unclear - no resource list or insufficient detail provided

6a. What additional resources are required for this station?

....

7. Has the station been designed to allow sufficient time for each task, and for all tasks to be completed?

○ Yes ○ No (go to 7a)

7a. Please select the most appropriate response;

□ Insufficient time has been allocated for all the tasks to be completed in this station.

□ Station requires deviation from best practice to complete task in time allowed e.g. student must take short cuts or other modifications to their usual approach to task. (Please describe) ....

□ Station has adequate time for all tasks to be completed; however, timing for individual tasks needs modification. (Please describe)..

8. Do the student or candidate instructions indicate how much time is permitted for each task?

○ Yes ○ No (please describe)

....

....

9. Does the order of tasks follow a logical sequence for this station?

○ Yes ○ No (please describe)

....

....

10. Does this station follow current clinical guidelines/ best practice for the task(s)?

○ Yes ○ No (please describe)

....

....

11. The inclusion of alcohol hand rub in all stations (including communication, history and examination/procedural) is necessary to reinforce this key patient safety behaviour. Does this station include hand washing/hand hygiene considerations?

○ Yes ○ No (please describe)

....

....

12. Do the student instructions provide clear information about the key tasks on which they will be assessed?

○ Yes ○ No (please describe)

....

....

13. Is there alignment between the task marks and the station objectives?

○ Yes ○ No (please describe)

....

....

14. Have any potential risks for student, patient or examiner been identified and managed?

○ Yes ○ No (please describe)

....

....

15. Health professionals need to adopt lifelong learning behaviours. Sometimes examinations cover material that is from previous years' curriculum to reinforce this message. Are the students being assessed on material they would have been expected to encounter in the current curriculum?

○ Yes ○ No ○ No, but this is being done intentionally. *We know that assessment drives learning.*

16. Is there anything in the format or expectations of this station that may have a negative impact on the students' clinical skills practice in the future?

○ Yes (go to 16a) ○ No ○ Maybe (go to 16a)

16a. Please describe your concerns the message this station will send to the student.

....

....

17. Examiners expect academic institutions to promote best practice in clinical skills. This may mean a station wording leads to unintended behavioural changes in the clinical practice of an examiner. Is there anything in the content of this station that may have a negative impact on an examiners' clinical practice following this OSCE?

○ Yes (go to 17a) ○ No ○ Maybe (go to 17a)

17a. Please describe your concerns the message this station will send to the student.

....

....

18. Is there anything in the content of this station that may be misleading to the candidate due to problems with authenticity? For example, would a patient really present in this way and does this presentation match any diagnosis a candidate is expected to identify through clinical reasoning?

○ Yes (please describe) ○ No

....

....

19. Have any other errors in this station been identified that have not been adequately covered by the preceding questions?

○ Yes (please describe) ○ No

....

....

20. YOUR SUMMARY... Based on the above evaluation this OSCE station requires:

○ Redevelopment (go to A)

○ Refinement (go to B)

○ Retirement (go to C)

○ Respect (go to D)

A. You have identified that this station requires major modifications.
Please outline all changes you would recommend to the station's author for quality improvement.

....

....

B. You have identified that this station requires minor changes. Please outline any recommended changes.

....

....

C. You have identified a station that you believe is beyond redemption or has reached the end of its usefulness.

D. Congratulations - you have uncovered a useable station

Thank you for using OWSAT as part of your OSCE quality improvement processes.

Appendix F: Rationale behind inclusion of each question in final version of OWSAT

Question	Explanation for inclusion in tool	Explanation included in tool
Please provide the following information.		
Name	Contact details for providing copy of feedback to station writers and reviewers.	Yes.
Email address		
OSCE Station Name	For station identification.	No.
Q 1. What is the source of this Station? Please tick all that apply.		
Preclinical undergraduate years	To assist reviewers to consider the context of the assessment and the level of training of the candidate.	No.
Early clinical undergraduate years		
Late or exit clinical undergraduate years		
Vocational training years		
Other		
Q 2. Is the author of this OSCE Station known?		
Unknown (1)	For feedback purposes. This key faculty development step should form one pillar of the quality improvement cycle.	No.
Known (2)		
Q 3. Please describe the task or tasks to be performed in this station.		
	The task or tasks should be identifiable by the reviewer as well as the candidate and examiner.	No.
Q 4. Has the appropriate assessment format been selected for each task in this station?		
Yes (1)	This question relates to resource issues where pure knowledge items are tested in the expensive OSCE format.	No.
No (2)		
Q 5. Are there adequate resources, equivalent to what would be found in normal clinical practice?		
Yes (1)	This question covers concerns about provision of the correct tools for the task to enable safe and authentic practice, e.g. gloves, sharps containers, full set of neurological equipment.	No.
No (2)		
Unclear - no resource list or insufficient detail provided (3)		
Q 6. Has the station been designed to allow sufficient time for each task, and for all tasks to be completed?		
Yes (1)	Insufficient time to perform tasks forces candidates to make choices about what to leave out or rush through tasks risking mistakes. Undermines patient-centred practice.	No, but risks are suggested in Q6a, reached by skip logic from Q 6 answer No as shown.
No (2)		
Q 6a. Please select the most appropriate response;		
Insufficient time has been allocated for all the tasks to be completed in this station. (1)		
Station requires deviation from best practice to complete task in time allowed e.g. student must take short cuts or other modifications to their usual approach to task. (Please describe) (2)		
Station has adequate time for all tasks to be completed; however, timing for individual tasks needs modification. (Please describe) (3)		
Q 7. Do the student or candidate instructions indicate how much time is permitted for each task		
Yes (1)	Candidate time management requires this knowledge.	No.
No (Please describe) (2) ________________________		
Q 8. Does the order of tasks follow a logical sequence for this station?		
Yes (1)	Candidate may become confused, may also affect SP performance standardisation.	No.
No (Please describe) (2)		
Q 9. Does this station follow current clinical guidelines/ best practice for the task(s)?		
Yes (1)	Station may discriminate against a good candidate who is more up to date than the marking sheet or examiner.	No.
No (Please describe) (2)		

Question	Explanation for inclusion in tool	Explanation included in tool
Q 10. Do the student/candidate instructions provide clear information about the key tasks on which they will be assessed?		
Yes (1)	Unclear or misleading information will prevent candidate from demonstrating their competence.	No.
No (Please describe) (2)		
Q11. Is there alignment between the task marks and the station objectives?		
Yes (1)	Basic assessment principle; no marks for an objective means it is not actually being assessed.	No.
No (Please describe) (2)		
Q 12. Have any potential risks for student, patient or examiner been identified and managed?		
Yes (1)	Duty of care in workplace.	No.
No (Please describe) (2) _		
Q 13. Are the students being assessed on material they would have been expected to encounter in the current curriculum?		
Yes (1)	Should have been picked up in the blueprint process. OK to reinforce lifelong learning by assessing from a previous year. You need to adequately sample from current curriculum. A waste of resources to mistakenly assess from curriculum that has not been covered.	Health professionals need to adopt life long learning behaviours. Sometimes examinations cover material that is from previous years' curriculum to reinforce this message.
No (2)		
No, but this is being done intentionally. (3)		
Q 14. Is there anything in the format or expectations of this station that may have a negative impact on the students' clinical skills practice in the future?		
Yes (1)	Behaviours such as rote learning check-lists, not washing hands, not using gloves when performing tasks because none were provided in OSCE - unintended consequences of station-writing decisions.	We know that assessment drives learning.
No (2)		
Maybe (3)		
Q15. Is there anything in the content of this station that may have a negative impact on an examiners' clinical practice following this OSCE?		
Yes (1)	Examiner observing university or specialist college expectations of station may adopt this as best practice in future, including ordering of unnecessary investigations.	Examiners expect academic institutions to promote best practice in clinical skills. This may mean a station wording leads to unintended behavioural changes in the clinical practice of an examiner.
No (2)		
Maybe (3)		
Q 16. Is there anything in the content of this station that may be misleading to the candidate due to problems with authenticity?		
No (1)	Clinical reasoning is undermined if the correct information is not provided. A patient scenario with incongruent content may force candidate to make choices between two diagnoses, one correct, one incorrect. Test of luck not skill.	For example, would a patient really present in this way and does this presentation match any diagnosis a candidate is expected to identify through clinical reasoning?
Yes (Please describe) (2)		
Q 17. Have any other errors in this station been identified that have not been adequately covered by the preceding questions?		
No (1)	Opportunity to collate potential future modifications to OWSAT.	No.
Yes (Please describe) (2) ______________________		
FINALLY...Based on the above evaluation this OSCE station requires:		
Redevelopment (1)		
Refinement (2)		
Retirement (3)		
Respect (4)		

Appendix G: Presentations and Publications OSCE Errors 2012-2014

Year	Title	Conference	Type	Co-Authors	Relevance
2012	Whole of school involvement in review of OSCE station wording to improve quality of assessment	Ottawa Conference on Assessment, Kuala Lumpur, Malaysia	Oral presentation	G. Somers, S. Bullock, B. Chapman	Intro to errors and quality improvement processes in OSCE
2013	The development of a classification system for item writing errors in OSCE to assist in pre-examination quality improvement opportunities.	APMEC 2013, Singapore, Singapore	Oral presentation	S. Bullock, L. Sweet, G. Somers, J. Black	Concept of errors in OSCE/ Pre-tool list created
2013	Special Issue: Abstracts of the 10th Asia Pacific Medical Education Conference (APMEC), National University of Singapore, Singapore, 16–20 January 2013	Medical Education ª 2013 Blackwell Publishing Ltd. MEDICAL EDUCATION 2013; 47 (Suppl. 2): 1–16	Publication (Abstract - Conference Proceedings)	S. Bullock, L. Sweet, G. Somers, J. Black	Concept of errors in OSCE/ Pre-tool list created
2013	A Tool to Reduce Bias from Flawed OSCE Items at the Writing and Reviewing Stage of OSCE Development	CCME 2013, Quebec, Canada	Oral presentation	L. Sweet, S. Bullock, G. Somers	First version of Tool launched in Canada
2013	A Tool to Reduce Bias from Flawed OSCE Items at the Writing and Reviewing Stage of OSCE Development	Special Issue: Abstracts of the Canadian Conference on Medical Education, 20-23 April 2013, Ottawa, Canada	Medical Education ª 2013 Blackwell Publishing Ltd. MEDICAL EDUCATION 2013; 47 (Suppl. 1): 1–16	L. Sweet, S. Bullock, G. Somers	First version of Tool launched in Canada and then abstract is published
2013	A tool to reduce bias from flawed OSCE items at the writing and reviewing stage of OSCE development	ANZAHPE Conference, Melbourne,	PeARL session	L. Sweet, S Bullock, G. Somers	Peer review process for tool, utility and acceptability explored locally/nationally
2014	What difference does a minute make? Lessons learned when the OSCE bell tolled too early.	APMEC 2014 Singapore, Singapore	Oral presentation	S. Bullock, L. Sweet, G. Somers, J. Black, M.Shuttleworth	Further presentation on the theme of Errors in OSCE
2014	What difference does a minute make? Lessons learned when the OSCE bell tolled too early.	Special Issue: Abstracts of the 11th Asia Pacific Medical Education Conference (APMEC), National University of Singapore, Singapore, 16–20 January 2014	Medical Education ª 2014 Blackwell Publishing Ltd. MEDICAL EDUCATION 2014; 48 (Suppl. 2): 1–16	S. Bullock, L. Sweet, G. Somers, J. Black, M.Shuttleworth	Also published in conference proceedings
2014	The alpha problem	APMEC 2014 Singapore, Singapore	Poster presentation	J. Black, K. Brotchie	Mathematical concept errors definitions mapped with OSCE theme
2014	The Objective Structured Clinical Examination (OSCE) – Identification of stations level flaws through the use of an OSCE item writing error detection tool.	Ottawa Conference on Assessment, Ottawa, Canada	Workshop	L. Sweet, S. Bullock, G. Somers	Peer review process for tool, utility and acceptability explored internationally

Reference List

Adamo, G. (2003). Simulated And Standardized Patients In Osces: Achievements And Challenges 1992-2003. *Medical Teacher, 25*(3), 262-270. doi: doi:10.1080/0142159031000100300

Akkerman, S F, Bronkhorst, L H, & Zitter, I. (2013). The Complexity Of Educational Design Research. *Quality And Quantity, 47*(1), 421-439. doi: 10.1080/10494821003790863

Albanese, M A, Mejicano, G, Anderson, W M, & Gruppen, L. (2010). Building A Competency-Based Curriculum: The Agony And The Ecstasy. *Advances In Health Sciences Education, 15*(3), 439-454. doi: 10.1007/s10459-008-9118-2

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington D.C.: The American Educational Research Association.

Anderson, T, & Shattuck, J. (2012). Design-Based Research: A Decade Of Progress In Education Research? *Educational Researcher, 41*(1), 16-25. doi: 10.3102/0013189x11428813

Andreatta, P B, Marzano, D A, & Curran, D S. (2011). Validity: What Does It Mean For Competency-Based Assessment In Obstetrics And Gynecology? *American Journal Of Obstetrics And Gynecology, 204*(5), 384 e381-386. doi: 10.1016/j.ajog.2011.01.061

Anonymous. (1876). Examinations And The Medical Council. *The Lancet, June 17, 1876*, 893-894.

Arora, S, & Sevdalis, N. (2008). Hospex And Concepts Of Simulation. *Journal Of The Royal Army Medical Corps, 154*(3), 202-205.

Auewarakul, C, Downing, S M, Praditsuwan, R, & Jaturatamrong, U. (2005). Item Analysis To Improve Reliability For An Internal Medicine Undergraduate OSCE. *Advances In Health Sciences Education, 10*, 105-113. doi: 10.1007/s10459-005-2315-3

Bamber, A R, Quince, T A, Barclay, S I G, Clark, J D A, Siklos, P W L, & Wood, D F. (2014). Medical Student Attitudes To The Autopsy And Its Utility In Medical Education: A Brief Qualitative Study At One UK Medical School. *Anatomical Sciences Education, 7*(2), 87-96. doi: 10.1002/ase.1384

Banja, J. (2010). The Normalization Of Deviance In Healthcare Delivery. *Business Horizons, 53*(2), 139. doi: 10.1016/j.bushor.2009.10.006

Bannan-Ritland, B. (2003). The Role Of Design In Research: The Integrative Learning Design Framework. *Educational Researcher, 32*(1), 21-24.

Barab, S, & Squire, K. (2004). Design-Based Research: Putting A Stake In The Ground. *The Journal Of The Learning Sciences, 13*(1), 1-14.

Barman, A. (2005). Critiques On The Objective Structured Clinical Examination. *Annals Of Academic Medicine Singapore, 34*(8), 478-482.

Barrows, H S. (1987). *Simulated (Standardized) Patients and Other Human Simulations*. Springfield, Ill.,: Health Sciences Consortium.

Barrows, H S. (1993). An Overview Of The Uses Of Standardized Patients For Teaching And Evaluating Clinical Skills. *Academic Medicine, 68*(6 June), 443-451.

Barry, M, Bradshaw, C, & Noonan, M. (2013). Improving The Content And Face Validity Of OSCE Assessment Marking Criteria On An Undergraduate Midwifery Programme: A Quality Initiative. *Nurse Education In Practice, 13*(5), 477-480. doi: 10.1016/j.nepr.2012.11.006

Beckman, T J, Ghosh, A K, Cook, D A, Erwin, P J, & Mandrekar, J N. (2004). How Reliable Are Assessments Of Clinical Teaching? *Journal Of General Internal Medicine, 19*(9), 971-977. doi: http://dx.doi.org/10.1111/j.1525-1497.2004.40066.x

Berendonk, C, Stalmeijer, R E, & Schuwirth, L W T. (2013). Expertise In Performance Assessment: Assessors' Perspectives. *Advances In Health Sciences Education, 18*(4), 559-571. doi: 10.1007/s10459-012-9392-x

Bland, J M, & Altman, D G. (1977). Statistical Notes: Cronbach's Alpha. *British Medical Journal, 314*(7080), 572.

Bleakley, A. (2003). ‘Good’ And ‘Poor’ Communication In An OSCE: Education Or Training? *Medical Education, 37*(3), 186-187. doi: 10.1046/j.1365-2923.2003.01423.x

Bok, H G J, Teunissen, P W, Favier, R P, Rietbroek, N J, Theyse, L F H, Brommer, H, Haarhuis, J C M, van Beukelen, P, Van der Vleuten, C P M, & Jaarsma, D A D C. (2013). Programmatic Assessment Of Competency-Based Workplace Learning: When Theory Meets Practice. *BMC Medical Education, 13*(123), 1-9.

Boulet, J, Ben-David, M F, Hambleton, R K, Burdick, W, Ziv, A, & Gary, N. (1998). An Investigation Of The Sources Of Measurement Error In The Post-Encounter Written Scores From Standardized Patient Examinations. *Advances In Health Sciences Education, 3*, 89-100.

Boulet, J R. (2008). Summative Assessment In Medicine: The Promise Of Simulation For High-Stakes Evaluation. *Academic Emergency Medicine, 15*(11), 1017-1024. doi: 10.1111/j.1553-2712.2008.00228.x

Boursicot, K (Producer). (2003, Thursday 24 March 2011). OSCE Train The Trainer Assessment Workshop. *Hong Kong International Consortium*.

Boursicot, K, Etheridge, L, Ker, J, Sambandam, E, Setna, Z, Smee, S, & Sturrock, A. (2010). *Ottawa 2010 Performance Assessment Theme Group report.* Paper presented at the Ottawa Conference

Boursicot, K, Etheridge, L, Setna, Z, Sturrock, A, Ker, J, Smee, S, & Sambandam, E. (2011). Performance In Assessment: Consensus Statement And Recommendations From The Ottawa Conference. *Medical Teacher, 33*(5), 370-383. doi: 10.3109/0142159X.2011.565831

Boursicot, K, & Roberts, T. (2005). How To Set Up An OSCE. *The Clinical Teacher, 2*(1), 16-20. doi: 10.1111/j.1743-498X.2005.00053.x

Boursicot, K, Roberts, T, & Pell, G. (2006). Standard Setting For Clinical Competence At Graduation From Medical School: A Comparison Of Passing Scores Across Five Medical Schools. *Advances In Health Sciences Education, 11*, 173-183. doi: 10.1007/s10459-005-5291-8

Boursicot, K A, Roberts, T E, & Pell, G. (2007). Using Borderline Methods To Compare Passing Standards For Osces At Graduation Across Three Medical Schools. *Medical Education, 41*(11), 1024-1031. doi: 10.1111/j.1365-2923.2007.02857.x

Bracken, D W, & Rose, D S. (2011). When Does 360-Degree Feedback Create Behavior Change? And How Would We Know It When It Does? *Journal Of Business And Psychology, 26*(2), 183-192. doi: 10.1007/s10869-011-9218-5

Brailovsky, C, & Grand'Maison, P. (2000). Using Evidence to Improve Evaluation: A Comprehensive Psychometric Assessment of a SP-Based OSCE Licensing Examination. *Advances In Health Sciences Education, 5*, 207-219.

Britt, H, Miller, G, Charles, J, Henderson, J, Valenti, L, Harrison, C, Zhang, C, Chambers, T, Pollack, A J, Bayram, C, O'Halloran, J, & Pan, Y. (2012). *A Decade Of Australian General Practice Activity 2002-03 To 2011-12 General Practice Series Number 32* (Pp. 178).

Broadfoot, P, & Black, P. (2004). Redefining Assessment? The First Ten Years Of Assessment In Education. *Assessment In Education: Principles, Policy & Practice, 11*(1), 7-26. doi: 10.1080/0969594042000208976

Broome, J. (1991). “Utility”. *Economics And Philosophy, 7*(01), 1-12.

Brosnan, M, Evans, W, Brosnan, E, & Brown, G. (2006). Implementing Objective Structured Clinical Skills Evaluation (OSCE) In Nurse Registration Programmes In A Centre In Ireland: A Utilisation Focused Evaluation. *Nurse Education Today, 26*(2), 115-122. doi: 10.1016/j.nedt.2005.08.003

Brotchie, K, Somers, G, Bullock, A, Chapman, B, & Sweet, L. (2012). *Whole Of School Involvement In Review Of OSCE Station Wording To Improve Quality Of Assessment.* Paper presented at the Ottawa Conference, Kuala Lumpur.

Brotchie, K, Sweet, L, Bullock, S, & Black, J. (2013). The Development Of A Classification System For Item Writing Errors In OSCE To Assist In Pre-Examination Quality Improvement Opportunities. Paper presented at the Asia Pacific Medical Education Conference Singapore.

Brown, A L. (1992). Design Experiments: Theoretical And Methodological Challenges In Creating Complex Interventions In Classroom Settings. *Journal of the Learning Sciences, 2*(2), 141-178. doi: 10.1207/s15327809jls0202_2

Brown, R, & Skinner, D. (2003). Setting Up The Membership Examination. *Emergency Medicine Journal, 20*(4 July (Supplement)), S1-2.

Casey, P M, Goepfert, A R, Espey, E L, Hammoud, M M, Kaczmarczyk, J M, Katz, N T, Neutens, J J, Nuthalapaty, F S, & Peskin, E. (2009). To The Point: Reviews In Medical Education-The Objective Structured Clinical Examination. *American Journal of Obstetrics & Gynecology, January*(January), 25-34. doi: 10.1016/j.ajog.2008.09.878

Chandratilake, M, Davis, M H, & Ponnamperuma, G. (2010). Evaluating And Designing Assessments For Medical Education: The Utility Formula. *The Internet Journal of Medical Education, 1*(1).

Chipman, J G, Beilman, G J, Schmitz, C C, & Seatter, S C. (2007). Development And Pilot Testing Of An OSCE For Difficult Conversations In Surgical Intensive Care. *Journal of Surgical Education, 64*(2), 79-87. doi: 10.1016/j.jsurg.2006.11.001

Christie, J, Pryor, E, & Paull, A M. (2011). Presenting Under Pressure: Communication And International Medical Graduates. *Medical Education, 45*(5), 532-532. doi: 10.1111/j.1365-2923.2011.03955.x

Clarke, C, Harcourt, M, & Flynn, M. (2012). Clinical Governance, Performance Appraisal And Interactional And Procedural Fairness At A New Zealand Public Hospital. *Journal Of Business Ethics, 117*(3), 667-678. doi: 10.1007/s10551-012-1550-9

Cohen, D, Colliver, J, Robb, R, & Swartz, M. (1997). A Large-Scale Study Of The Reliabilities Of Checklist Scores And Ratings Of Interpersonal And Communication Skills Evaluated On A Standardized-Patient Examination. *Advances In Health Sciences Education, 1*, 209-213.

Cohen, R, Reznick, R, Taylor, B, Provan, J, & Rothman, A. (1990). Reliability And Validity Of The Objective Structured Clinical Examination In Assessing Surgical Residents. *The Americal Journal Of Surgery, 160*(September), 302-305.

Cook, D A. (2014). When I Say… Validity. *Medical Education, 48*(10), 948-949. doi: 10.1111/medu.12401

Cook, D A, & Beckman, T J. (2006). Current Concepts In Validity And Reliability For Psychometric Instruments: Theory And Application. *American Journal Of Medicine, 119*(2), 166 e167-116. doi: 10.1016/j.amjmed.2005.10.036

Course, F. (2012). Fundamentals Of Assessment In Medical Education. Retrieved 21 September, 2014, from http://www.famecourse.org/

Cramer, J A, Roy, A, Burrell, A, Fairchild, C J, Fuldeore, M J, Ollendorf, D A, & Wong, P K. (2008). Medication Compliance And Persistence: Terminology And Definitions. *Value Health, 11*(1), 44-47. doi: 10.1111/j.1524-4733.2007.00213.x

Crichton, N J. (1998). Statistical Considerations In Design And Analysis. In B. Roe & C. Webb (Eds.), *Research And Development In Nursing Practice* (pp. 190-215). Whurr, London: Wiley Interscience.

Cunnington, J, Neville, A, & Norman, G. (1997). The Risks Of Thoroughness: Reliability And Validity Of Global Ratings And Checklists In An OSCE. *Advances In Health Sciences Education, 1*, 227-233.

Curtis, M, McNaughton, N, Robb, A, & Tabak, D. (1994). How To Run An OSCE [Video]. Toronto: Department of Family and Community Medicine, University of Toronto.

De Champlain, A F. (2010). A Primer On Classical Test Theory And Item Response Theory For Assessments In Medical Education. *Medical Education, 44*(1), 109-117. doi: 10.1111/j.1365-2923.2009.03425.x

Dede, C. (2005). Why Design-Based Research is Both Important and Difficult. *Educational Technology, 45*(1 (January-February)), 5-8.

Dijksterhuis, M, Schuwirth, L, Braat, D, & Scheele, F. (2011). What's The Problem With The Mini-CEX? *Medical Education, 45*(3), 318-319. doi: 10.1111/j.1365-2923.2010.03927.x

Dolmans, D H J M, & Tigelaar, D. (2012). Building Bridges Between Theory And Practice In Medical Education Using A Design-Based Research Approach: AMEE Guide No. 60. *Medical Teacher, 34*(1), 1-10. doi: 10.3109/0142159X.2011.595437

Dolmans, D H J M, & van der Vleuten, C P M. (2010). Research In Medical Education: Practical Impact On Medical Training And Future Challenges. *Netherlands Journal Of Medical Education, 29*(1), 3-9.

Donaldson, L. (2009). WHO Patient Safety Curriculum Guide for Medical Schools: WHO Library Cataloguing-in-Publication Data.

Donaldson, L, Appleby, L, Boyce, J, Buckley, M, Drife, J, Firth-Cozens, J, Hart, P, Kirkup, B, de Leval, M, Naftalin, N, Reason, J, Rigge, M, Smart, K, Toft, B, Vincent, C, Walker, S, Williams, J L, & Worth, B. (2000). *An Organisation with a Memory*. London, UK: The Stationery Office.

Dornan, T, Hadfield, J, Brown, M, Boshuizen, H, & Scherpbier, A. (2005). How Can Medical Students Learn In A Self-Directed Way In The Clinical Environment? Design-Based Research. *Medical Education, 39*(4), 356-364. doi: 10.1111/j.1365-2929.2005.02112.x

Downing, S M. (2003). Validity: On The Meaningful Interpretation Of Assessment Data. *Medical Education, 37*, 830-837.

Downing, S M. (2004). Reliability: On The Reproducibility Of Assessment Data. *Medical Education, 38*, 1006-1012. doi: 10.1046/j.1365-2929.2004.01932.x

Dreyfus, S E, & Dreyfus, H L. (1980). *A Five-Stage Model Of The Mental Activities Involved In Directed Skill Acquisition*. Research Paper. University of California. Berkeley. Operations Research Centre. Retrieved from https://www.google.com.au/webhp?sourceid=chrome-instant&rlz=1C5CHFA_enAU529AU535&ion=1&espv=2&ie=UTF-8#q=Dreyfus+S%2C+Dreyfus+H.+1980.+A+five+stage+model+of+the+mental+activities+involved+in+directed+skill+acquisition%2C+Research+Paper%2C+California+University+Berkeley+Operations+Research+Center%2C+A155480.

Eaton, D M, & Cottrell, D. (1999). Structured Teaching Methods Enhance Skill Acquisition But Not Problem-Solving Abilities: An Evaluation Of The 'Silent Run Through'. *Medical Education, 33*, 019-023.

Eberhard, L, Hassel, A, Bäumer, A, Becker, F, Beck-Mußotter, J, Bömicke, W, Corcodel, N, Cosgarea, R, Eiffler, C, Giannakopoulos, N N, Kraus, T, Mahabadi, J, Rues, S, Schmitter, M, Wolff, D, & Wege, K C. (2011). Analysis Of Quality And Feasibility Of An Objective Structured Clinical Examination (OSCE) In Preclinical Dental Education. *European Journal Of Dental Education, 15*(3), 172-178. doi: 10.1111/j.1600-0579.2010.00653.x

Edler, A A, & Fanning, R M. (2007). "A Rose By Any Other Name"? Toward A Common Terminology In Simulation Education And Assessment. *Critical Care Medicine, 35*(9), 2237-2238; author reply 2238. doi: 10.1097/01.CCM.0000281643.88046.DC

Educause. (2012). 7 Things You Should Know About Educational Design Research. In Educause Learning Initiative (Ed.).

Eitel, F, Kanz, K-G, & Tesche, A. (2000). Training And Certification Of Teachers And Trainers: The Professionalism Of Medical Education. *Medical Teacher, 22*(5), 517-526.

Ellaway, R H, Pusic, M, Yavner, S, & Kalet, A L. (2014). Context Matters: Emergent Variability In An Effectiveness Trial Of Online Teaching Modules. *Medical Education, 48*(4), 386-396. doi: 10.1111/medu.12389

Epstein, R M. (2007). Assessment in Medical Education. *The New England Journal Of Medicine, 356*, 387-396.

Epstein, R M, & Hundert, E M. (2002). Defining and Assessing Professional Competence. *Journal Of The American Medical Association, 287*(No. 2 (Reprinted)), 226-235.

Esmail, A, & Roberts, C. (2013). Academic Performance Of Ethnic Minority Candidates And Discrimination In The MRCGP Examinations Between 2010 And 2012: Analysis Of Data. *British Medical Journal, 347*, f5662. doi: 10.1136/bmj.f5662

Eva, K W, & Regehr, G. (2007). Knowing When to Look It Up: A New Conception of Self-Assessment Ability. *Academic Medicine, 82*(10 Suppl), S81-S84.

Eva, K W, & Regehr, G. (2011). Exploring The Divergence Between Self-Assessment And Self-Monitoring. *Advances In Health Sciences Education, 16*(3), 311-329. doi: 10.1007/s10459-010-9263-2

Fairhurst, K, Strickland, A, & Maddern, G J. (2011). Simulation Speak. *Journal of Surgical Education, 68*(5), 382-386. doi: 10.1016/j.jsurg.2011.03.003

Fleming, P R, Manderson, W G, Matthews, M B, Sanderson, P H, & Stokes, J F. (1974). Evolution Of An Examination: M.R.C.P. (U.K.). *British Medical Journal, 2*(5910), 99-107.

Fromme, H B, Karani, R, & Downing, S M. (2009). Direct Observation In Medical Education: A Review Of The Literature And Evidence For Validity. *Mount Sinai Journal Of Medicine, 76*(4), 365-371. doi: 10.1002/msj.20123

Fuller, R, Homer, M, & Pell, G. (2013). Longitudinal Interrelationships Of OSCE Station Level Analyses, Quality Improvement And Overall Reliability. *Medical Teacher, 35*(6), 515-517. doi: 10.3109/0142159X.2013.775415

Gawron, V J, Drury, C G, Fairbanks, R J, & Berger, R C. (2006). Human Factors Engineering: Where Are We Now? *American Journal Of Medical Quality, 21*, 57-67.

Given, L M. (2008). *The SAGE Encyclopedia of Qualtitative Research Methods* (Vol. 1& 2). California: SAGE Publications, Inc.

Gonczi, A. (1994). Competency Based Assessment In The Professions In Australia. *Assessment In Education: Principles, Policy & Practice, 1*(1), 27-44. doi: 10.1080/0969594940010103

Gormley, G, Sterling, M, Menary, A, & McKeown, G. (2012). Keeping It Real! Enhancing Realism In Standardised Patient OSCE Stations. *The Clinical Teacher, 9*, 382-386.

Gorsira, M. (2009). The Utility Of (European) Licensing Examinations. AMEE Symposium, Prague 2008. *Medical Teacher, 31*(3), 221-222.

Guba, E G, & Lincoln, Y S. (1988). Do Inquiry Paradigms Imply Inquiry Methodologies? In D. M. Fetterman (Ed.), *Qualitative Approaches To Evaluation In Education. The Silent Scientific Revolution.* (pp. 89-115). New York: Praeger Publishers.

Gupta, P, Dewan, P, & Singh, T. (2010). Objective Structured Clinical Examination (OSCE) Revisited. *Indian Pediatrics, 47*, 911-920.

Hager, P, Gonczi, A, & Athanasou, J. (1994). General Issues About Assessment Of Competence. *Assessment & Evaluation In Higher Education, 19*(1), 3-16. doi: 10.1080/0260293940190101

Haidet, Paul, Kelly, Adam, Chou, Calvin, & The Communication, Curriculum, and Culture Study Group,. (2005). Characterizing The Patient-Centeredness Of Hidden Curricula In Medical Schools: Development And Validation Of A New Measure. *Academic Medicine, 80*, 44-50.

Halperin, E C. (2011). Abraham Flexner And The Evolution Of The Modern Medical School. *Medical Education, 45*(1), 10-12. doi: 10.1111/j.1365-2923.2010.03899.x

Harden, R, Crosby, J, Davis, M H, & Friedman, M. (1999). AMEE Guide No. 14: Outcome-Based Education: Part 5 - From Competency To Meta-Competency: A Model For The Specification Of Learning Outcomes. *Medical Teacher, 21*(6), 546-552.

Harden, R M. (1979a). How to... Assess Clinical Competence - An Overview. *Medical Teacher, 1*(6), 289-296.

Harden, R M. (1979b). The M.R.C.G.P. Examination. *The Lancet, 314*(8138), 367. doi: 10.1016/s0140-6736(79)90386-6

Harden, R M. (1985). Assessment Of Clinical Competence Examiners' Tool-Kit. In I. R. Hart, R. M. Harden & H. J. Walton (Eds.), *Newer Developments In Assessing Clinical Competence* (pp. 11-19). Montreal: Heal Publications.

Harden, R M. (1988). What Is An OSCE? *Medical Teacher, 10*(1), 19-22.

Harden, R M, & Gleeson, F A. (1979). Assessment Of Clinical Competence Using An Objective Structured Clinical Examination (OSCE). ASME Booklet No. 8. *Medical Education, 13*, 44-54.

Harden, R M, Stevenson, M, Wilson Downie, W, & Wilson, G M. (1975). Assessment Of Clinical Competence Using Objective Structured Examination. *British Medical Journal, 1*(5955 22 February 1975), 447-451.

Hattie, J, Jaeger, R M, & Bond, L. (1999). Persistent Methodological Questions in Educational Testing *Review Of Research in Education, 24* (Vol. 24, pp. 393-446).

Hawkins, R E, Margolis, M J, Durning, S J, & Norcini, J J. (2010). Constructing A Validity Argument For The Mini-Clinical Evaluation Exercise: A Review Of The Research. *Academic Medicine, 85*(9), 1453-1461. doi: 10.1097/ACM.0b013e3181eac3e6

Hays, R. (2008). Assessment In Medical Education: Roles For Clinical Teachers. *The Clinical Teacher, 5*, 23-27.

Herrington, J, McKenney, S, Reeves, T, & Oliver, R. (2007). *Design-Based Research And Doctoral Students: Guidelines For Preparing A Dissertation Proposal.* Paper presented at the Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, Chesapeake, VA.

Hettinga, A M, Denessen, E, & Postma, C T. (2010). Checking The Checklist: A Content Analysis Of Expert- And Evidence-Based Case-Specific Checklist Items. *Medical Education, 44*(9), 874-883. doi: 10.1111/j.1365-2923.2010.03721.x

Hingle, S T, Robinson, S, Colliver, J A, Rosher, R B, & McCann-Stone, N. (2011). Systems-Based Practice Assessed With A Performance-Based Examination Simulated And Scored By Standardized Participants In The Health Care System: Feasibility And Psychometric Properties. *Teaching And Learning In Medicine, 23*(2), 148-154. doi: 10.1080/10401334.2011.561751

Hodges, B. (2003a). Osce! Variations On A Theme By Harden. *Medical Education, 37*, 1134-1140. doi: 10.1046/j.1365-2923.2003.01717.x

Hodges, B. (2003b). Validity And The OSCE. *Medical Teacher, 25*(3), 250-254. doi: 10.1080/01421590310001002836

Hodges, B. (2009). *The Objective Structured Clinical Examination: A Socio-History* (First ed.): LAP Lambert Academic Publishing.

Hodges, B. (2012). *The Future Of Assessment: Learning To Love The Collective And The Subjective*. Paper presented at the 15th Ottawa International Conference on Clinical Assessment, Kuala Lumpur.

Hodges, B D. (2007). A Socio-Historical Study Of The Birth And Adoption Of The Objective Structured Clinical Examination (OSCE). (Doctor of Philosophy), University of Toronto, Toronto, Canada.

Holmboe, E S, Ward, D S, Reznick, R K, Katsufrakis, P J, Leslie, K M, Patel, V L, Ray, D D, & Nelson, E A. (2011). Faculty Development In Assessment: The Missing Link In Competency-Based Medical Education. *Academic Medicine, 86*(4), 460-467. doi: 10.1097/ACM.0b013e31820cb2a7

Hopper, K. (2008). Qualitative And Quantitative Research: Two Cultures. *Psychiatric Services, 59*(7), 711.

Hu, W C Y, McColl, G J, Thistlethwaite, J E, Schuwirth, L W T, & Wilkinson, T. (2013). Where Is The Next Generation Of Medical Educators? *Medical Journal Of Australia, 198*(1), 8-9.

Hurley, K. (2005). *OSCE And Clinical Skills Handbook*. Toronto: Elsevier/Saunders.

Ilgen, J S, Bowen, J L, McIntyre, L A, Banh, K V, Barnes, D, Coates, W C, Druck, J, Fix, M L, Rimple, D, Yarris, L M, & Eva, K W. (2013). Comparing Diagnostic Performance And The Utility Of Clinical Vignette-Based Assessment Under Testing Conditions Designed To Encourage Either Automatic Or Analytic Thought. *Academic Medicine, 88*(10), 1545-1551. doi: 10.1097/ACM.0b013e3182a31c1e

Ilic, D, Nordin, R, Glasziou, P, Tilson, J, & Villaneuva, E. (2014). Development And Validation Of The ACE Tool: Assessing Medical Trainees' Competency In Evidence Based Medicine. *BMC Medical Education, 14*(114), 1-6. doi: 10.1186/1472-6920-14-114

Iramaneerat, C, Yudkowsky, R, Myford, C M, & Downing, S M. (2008). Quality Control Of An OSCE Using Generalizability Theory And Many-Faceted Rasch Measurement. *Advances In Health Sciences Education, 13*(4), 479-493. doi: 10.1007/s10459-007-9060-8

Jansen, J, Tan, L, van der Vleuten, C, van Luijk, S J, Rethans, J J, & Grol, R. (1995). Assessment Of Competence In Technical Clinical Skills Of General Practitioners. *Medical Education, 29*, 247-253.

Kan Ma, H, Min, C, Neville, A, & Eva, K. (2013). How Good Is Good? Students and Assessors’ Perceptions of Qualitative Markers of Performance. *Teaching And Learning In Medicine, 25*(1), 15-23. doi: 10.1080/10401334.2012.741545

Kane, M T. (1992). An Argument-Based Approach to Validity. *Psychological Bulletin, 112*(3), 527-535.

Kane, M T. (2001). Current Concerns In Validity Theory. *Journal Of Educational Measurement, 38*(Winter 2001, No. 4), 319-342.

Kane, T. (2014). Whose Lingua Franca?: The Politics of Language in Transnational Medical Education. *The Journal Of General Education, 63*(2-3), 94-112. doi: 10.1353/jge.2014.0015

Kao, D T. (2013). The Impacts Of Goal Orientation, Terminology Effect, And Source Credibility On Communication Effectiveness. *Journal of Applied Social Psychology, 43*(10), 2007-2016. doi: 10.1111/jasp.12154

Kaslow, N J, Rubin, N J, Bebeau, M J, Leigh, I W, Lichtenberg, J W, Nelson, P D, Portnoy, S M, & Smith, I L. (2007). Guiding Principles And Recommendations For The Assessment Of Competence. *Professional Psychology: Research and Practice, 38*(5), 441-451. doi: 10.1037/0735-7028.38.5.441

Kaufman, R, & Keller, J M. (1994). Levels Of Evaluation: Beyond Kirkpatrick. *Human Resource Development Quarterly, 5*(4 Winter), 371-380.

Kelly, A. (2004). Design Research In Education: Yes, But Is It Methodological? *Journal Of The Learning Sciences, 13*(1), 115-128. doi: 10.1207/s15327809jls1301_6

Khan, K, & Ramachandran, S. (2012). Conceptual Framework For Performance Assessment: Competency, Competence And Performance In The Context Of Assessments In Healthcare – Deciphering The Terminology. *Medical Teacher, 34*(11), 920-928. doi: 10.3109/0142159X.2012.722707

Khan, K Z, Gaunt, K, Ramachandran, S, & Pushkar, P. (2013a). The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part II: Organisation & Administration. *Medical Teacher, 35*(9), e1447-1463. doi: 10.3109/0142159X.2013.818635

Khan, K Z, Ramachandran, S, Gaunt, K, & Pushkar, P. (2013b). The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: An Historical And Theoretical Perspective. *Medical Teacher, 35*(9), e1437-1446. doi: 10.3109/0142159X.2013.818634

King, A M, Perkowski‐Rogers, L C, & Pohl, H S. (1994). Planning Standardized Patient Programs: Case Development, Patient Training, And Costs. *Teaching and Learning in Medicine, 6*(1), 6-14. doi: 10.1080/10401339409539636

Klass, D J. (1994). “High‐Stakes”; Testing Of Medical Students Using Standardized Patients. *Teaching And Learning In Medicine, 6*(1), 28-32. doi: 10.1080/10401339409539639

Klein, G. (2008). Naturalistic Decision Making. *Human Factors, 50*(3), 456-460. doi: 10.1518/001872008x288385

Kogan, J R, Conforti, L, Bernabeo, E, Iobst, W, & Holmboe, E. (2011). Opening The Black Box Of Clinical Skills Assessment Via Observation: A Conceptual Model. *Medical Education, 45*(10), 1048-1060. doi: 10.1111/j.1365-2923.2011.04025.x

Kohn, L T, Corrigan, J M, Donaldson, M S, Committee on Quality of Health Care in America, & Institute of Medicine. (2000). To Err Is Human: Building A Safer Health System. In L. T. Kohn, J. M. Corrigan & M. S. Donaldson (Eds.), (pp. 312). National Academies Press: Institute of Medicine.

Kruger, J, & Dunning, D. (1999). Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments. *Journal Of Personality And Social Psychology, 77*(6), 1121-1134.

Levine, A I, Schwartz, A D, Bryson, E O, & Demaria, S, Jr. (2012). Role Of Simulation In U.S. Physician Licensure And Certification. *Mount Sinai Journal Of Medicine, 79*(1), 140-153. doi: 10.1002/msj.21291

Liabsuetrakul, T, Sirirak, T, Boonyapipat, S, & Pornsawat, P. (2013). Effect Of Continuous Education For Evidence-Based Medicine Practice On Knowledge, Attitudes And Skills Of Medical Students. *Journal Of Evaluation In Clinical Practice, 19*(4), 607-611. doi: 10.1111/j.1365-2753.2012.01828.x

Liao, S, Hunt, E, & Chen, W. (2010). Comparison Between Inter-Rater Reliability And Inter-Rater Agreement In Performance Assessment. *Annals Of Academic Medicine Singapore, 39*, 613-618.

Maatsch, J. (1981). Assessment Of Clinical Competence On The Emergency Medicine Specialty Certification Examination: The Validity Of Examiner Ratings Of Simulated Clinical Encounters. *Annals Of Emergency Medicine, 10*(10), 504-507.

McKenney, S, & Reeves, T C. (2013). Systematic Review Of Design-Based Research Progress: Is A Little Knowledge A Dangerous Thing? *Educational Researcher, 42*(2), 97-100. doi: 10.3102/0013189x12463781

McKinley, R K, Strand, J, Gray, T, Schuwirth, L, Alun-Jones, T, & Miller, H. (2008). Development Of A Tool To Support Holistic Generic Assessment Of Clinical Procedural Skills. *Medical Education, 42*, 619-627. doi: 10.1111/j.1365-2923.2008.03023.x

McManus, I C, & Wakeford, R. (2014). PLAB And UK Graduates' Performance On MRCP(UK) And MRCGP Examinations: Data Linkage Study. *British Medical Journal, 348*, g2621. doi: 10.1136/bmj.g2621

Memon, M A, Joughin, G R, & Memon, B. (2010). Oral Assessment And Postgraduate Medical Examinations: Establishing Conditions For Validity, Reliability And Fairness. *Advances In Health Sciences Education, 15*(2), 277-289. doi: 10.1007/s10459-008-9111-9

Messick, S. (1989). Meaning And Values In Test Validation: The Science And Ethics Of Assessment. *Educational Researcher, 18*(2), 5-11.

Messick, S. (1995). Validity Of Psychological Assessment. *American Psychologist*, 741-749.

Miller, G E. (1990). The Assessment Of Clinical Skills/ Competence/ Performance. *Academic Medicine, 65*(Number 9 September Supplement), S63-67.

Mitchell, M L, Henderson, A, Groves, M, Dalton, M, & Nulty, D. (2009). The Objective Structured Clinical Examination (OSCE): Optimising Its Value In The Undergraduate Nursing Curriculum. *Nurse Education Today, 29*(4), 398-404. doi: 10.1016/j.nedt.2008.10.007

Morse, J M. (2012). *Qualitative Health Research: Creating A New Discipline* Retrieved from http://MONASH.eblib.com.au/patron/FullRecord.aspx?p=1017563

Naeem, N, van der Vleuten, C, & Alfaris, E A. (2012). Faculty Development On Item Writing Substantially Improves Item Quality. *Advances In Health Sciences Education, 17*(3), 369-376. doi: 10.1007/s10459-011-9315-2

National Highway Traffic Safety Administration. (2005). The National EMS Scope Of Practice Model. In U.S. Department Of Transportation (Ed.). Washington, D.C.: National Highway Traffic Safety Administration.

Nestel, D. (2013). A Global Perspective On Postgraduate Medical Education. *Journal Of Health Specialties, 1*(1), 49. doi: 10.4103/1658-600x.110675

Newble, D. (2004). Techniques For Measuring Clinical Competence: Objective Structured Clinical Examinations. *Medical Education, 38*, 199-203. doi: 10.1046/j.1365-2923.2004.01755.x

Newble, D I, & Jaeger, K. (1983). The Effect Of Assessments And Examinations On The Learning Of Medical Students. *Medical Education, 17*, 165-171.

Newble, D J, & Swanson, D B. (1988). Psychometric Characteristics Of The Objective Structured Clinical Examination. *Medical Education, 22*, 325-334.

Noble, D J, & Donaldson, L J. (2011). Republished Paper: The Quest To Eliminate Intrathecal Vincristine Errors: A 40-Year Journey. *Postgraduate Medical Journal, 87*(1023), 71-74. doi: 10.1136/qshc.2008.030874rep

Norcini, J, Anderson, B, Bollela, V, Burch, V, Costa, M J, Duvivier, R, Galbraith, R, Hays, R, Kent, A, Perrott, V, & Roberts, T. (2011). Criteria For Good Assessment: Consensus Statement And Recommendations From The Ottawa 2010 Conference. *Medical Teacher, 33*(3), 206-214. doi: 10.3109/0142159X.2011.551559

Norcini, J, Boulet, J R, Dauphinee, W D, Opalek, A, Krantz, I, & Anderson, S T. (2010). Evaluating The Quality Of Care Provided By Graduates Of International Medical Schools. *Health Affairs, 29*(8), 1461-1468. doi: 10.1377/hlthaff.2009.0222

Norcini, J J, & Banda, S S. (2011). Increasing The Quality And Capacity Of Education: The Challenge For The 21st Century. *Medical Education, 45*(1), 81-86. doi: 10.1111/j.1365-2923.2010.03738.x

Norcini, J J, & McKinley, D W. (2007). Assessment Methods In Medical Education. *Teaching and Teacher Education, 23*(3), 239-250. doi: 10.1016/j.tate.2006.12.021

Norman, G. (2014). When I Say … Reliability. *Medical Education, 48*(10), 946-947. doi: 10.1111/medu.12511

Nulty, D D, Mitchell, M L, Jeffrey, C A, Henderson, A, & Groves, M. (2011). Best Practice Guidelines For Use Of Osces: Maximising Value For Student Learning. *Nurse Education Today, 31*(2), 145-151. doi: 10.1016/j.nedt.2010.05.006

Overeem, K, Wollersheim, H C, Arah, O A, Cruijsberg, J K, Grol, R, & Lombarts, K. (2012). Evaluation Of Physicians' Professional Performance: An Iterative Development And Validation Study Of Multisource Feedback Instruments. *BMC Health Services Research, 12*, 80. doi: 10.1186/1472-6963-12-80

Oxford Dictionaries. Retrieved June 24, 2015, from http://www.oxforddictionaries.com/definition/learner/feasibility

Parry, G, Cline, A, & Goldmann, D. (2012). Deciphering Harm Measurement. *Journal Of The American Medical Association, 307*(20), 2155-2156.

Patton, M Q. (1994). Developmental Evaluation. *Evaluation Practice, 15*(3), 311-319.

Pell, G, Fuller, R, Homer, M, & Roberts, T. (2013). Advancing The Objective Structured Clinical Examination: Sequential Testing In Theory And Practice. *Medical Education, 47*(6), 569-577. doi: 10.1111/medu.12136

Piryani, R M, Shankar, P R, Thapa, T P, Karki, B M, Kafle, R K, Khakurel, M P, & Bhandary, S. (2013). Introduction Of Structured Physical Examination Skills To Second Year Undergraduate Medical Students. *F1000Res, 2*, 16. doi: 10.12688/f1000research.2-16.v1

Poenaru, D, Morales, D, Richards, A, & O'Connor, M. (1997). Running An Objective Structured Clinical Examination On A Shoestring Budget. *The American Journal Of Surgery, 173*(June 1997), 538-541.

Polit, D F, & Beck, C T. (2006). The Content Validity Index: Are You Sure You Know What's Being Reported? Critique And Recommendations. *Research In Nursing And Health, 29*(5), 489-497. doi: 10.1002/nur.20147

Popham, W J. (2009). Assessment Literacy for Teachers: Faddish or Fundamental? *Theory Into Practice, 48*(1), 4-11. doi: 10.1080/00405840802577536

Popham, W J. (2011). Assessment Literacy Overlooked: A Teacher Educator's Confession. *The Teacher Educator, 46*(4), 265-273. doi: 10.1080/08878730.2011.605048

Prideaux, D. (2002). Research In Medical Education: Asking The Right Questions. *Medical Education, 36*, 1114-1115.

Quitter, S M. (1999). Assessment Literacy For Teachers: Making A Case For The Study Of Test Validity. *The Teacher Educator, 34*(4), 235-243. doi: 10.1080/08878739909555204

Reason, J. (2000). Human Error: Models And Management. *British Medical Journal, 320*(18 March 2000), 768-770.

Reeves, T C, McKenney, S, & Herrington, J. (2011). Publishing And Perishing: The Critical Importance Of Educational Design Research. *Australian Journal Of Educational Technology, 27*(1), 55-65.

Reznick, R, Blackmore, D, Cohen, R, Baumber, J, Rothman, A, Smee, S, Chalmers, A, Poldre, P, Birtwhistte, R, Walsh, P, Spady, D, & Berard, M. (1993a). Use Of Standardized-Patient Examinations In Conjunction With Licensure And

Certification. An Objective Structured Clinical Examination For The Licentiate Of The Medical Council Of Canada: From Research to Reality. *Academic Medicine, 68*(10 October Supplement), S4-S6.

Reznick, R, K., Smee, S, Baumber, J S, Cohen, R, Rothman, A, Blackmore, D, & Berard, M. (1993b). Guidelines For Estimating The Real Cost Of An Objective Structured Clinical Examination. *Academic Medicine, 68*(Number 7 July 1993), 513-517.

Ringsted, C, Hodges, B, & Scherpbier, A. (2011). 'The Research Compass': An Introduction To Research In Medical Education: AMEE Guide No. 56. *Medical Teacher, 33*(9), 695-709. doi: 10.3109/0142159X.2011.595436

Ringsted, C, Ostergaard, D, Ravn, L, Pedersen, J A, Berlac, P A, & van der Vleuten, C P M. (2003). A Feasibility Study Comparing Checklists And Global Rating Forms To Assess Resident Performance In Clinical Skills. *Medical Teacher, 25*(6), 654-658. doi: 10.1080/01421590310001605642

Roberts, C, Newble, D, Jolly, B, Reed, M, & Hampton, K. (2006). Assuring The Quality Of High-Stakes Undergraduate Assessments Of Clinical Competence. *Medical Teacher, 28*(6), 535-543. doi: 10.1080/01421590600711187

Roberts, C, Wass, V, Jones, R, Sarangi, S, & Gillett, A. (2003). A Discourse Analysis Study Of 'Good' And 'Poor' Communication In An OSCE: A Proposed New Framework For Teaching Students. *Medical Education, 37*(3), 192-201. doi: 10.1046/j.1365-2923.2003.01443.x

Sauer, J, Hodges, B, Santhouse, A, & Blackwood, N. (2005). The OSCE Has Landed: One Small Step for British Psychiatry? *Academic Psychiatry, 29*(3), 310-315.

Schneid, S D, Armour, C, Park, Y S, Yudkowsky, R, & Bordage, G. (2014). Reducing The Number Of Options On Multiple-Choice Questions: Response Time, Psychometrics And Standard Setting. *Medical Education, 48*(10), 1020-1027. doi: 10.1111/medu.12525

Schoonheim-Klein, M, Muijtjens, A, Habets, L, Manogue, M, van der Vleuten, C, & van der Velden, U. (2009). Who Will Pass The Dental OSCE? Comparison Of The Angoff And The Borderline Regression Standard Setting Methods. *European Journal Of Dental Education, 13*(3), 162-171. doi: 10.1111/j.1600-0579.2008.00568.x

Schou, L, Hostrup, H, Lyngso, E E, Larsen, S, & Poulsen, I. (2012). Validation Of A New Assessment Tool For Qualitative Research Articles. *Journal Of Advanced Nursing, 68*(9), 2086-2094. doi: 10.1111/j.1365-2648.2011.05898.x

Schuiling, K D, & Slager, J. (2000). Scope Of Practice: Freedom Within Limits. *Journal of Midwifery & Women's Health, 45*(6), 465-471.

Schultz, J-H, Nikendei, C, Weyrich, P, Möltner, A, Fischer, M R, & Jünger, J. (2008). Qualitätssicherung von Prüfungen am Beispiel des OSCE-Prüfungsformats: Erfahrungen der Medizinischen Fakultät der Universität Heidelberg. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen, 102*(10), 668-672. doi: 10.1016/j.zefq.2008.11.024

Schuwirth, L, Colliver, J, Gruppen, L, Kreiter, C, Mennin, S, Onishi, H, Pangaro, L, Ringsted, C, Swanson, D, van der Vleuten, C, & Wagner-Menghin, M. (2011). Research In Assessment: Consensus Statement And Recommendations From The Ottawa 2010 Conference. *Medical Teacher, 33*(3), 224-233. doi: 10.3109/0142159X.2011.551558

Schuwirth, L, & van der Vleuten, C. (2004). Merging Views On Assessment. *Medical Education, 38*, 1208-1211. doi: 10.1111/j.1365-2929.2004.02055

Schuwirth, L W, & van der Vleuten, C P. (2006). A Plea For New Psychometric Models In Educational Assessment. *Medical Education, 40*(4), 296-300. doi: 10.1111/j.1365-2929.2006.02405.x

Schuwirth, L W, & van der Vleuten, C P. (2011a). General Overview Of The Theories Used In Assessment: AMEE Guide No. 57. *Medical Teacher, 33*(10), 783-797. doi: 10.3109/0142159X.2011.611022

Schuwirth, L W, & van der Vleuten, C P. (2011b). Programmatic Assessment: From Assessment Of Learning To Assessment For Learning. *Medical Teacher, 33*(6), 478-485. doi: 10.3109/0142159X.2011.565828

Schuwirth, L W T, Southgate, L, Page, G G, Paget, N S, Lescop, J M J, Lew, S R, Wade, W B, & Baron-Maldonado, M. (2002). When Enough Is Enough: A Conceptual Basis For Fair And Defensible Practice Performance Assessment. *Medical Education, 36*, 925-930.

Schuwirth, L W T, & van der Vleuten, C P M. (2003). The Use Of Clinical Simulations In Assessment. *Medical Education, 37*(Suppl.1), 65-71.

Schuwirth, L W T, & van der Vleuten, C P M. (2013). How to design a useful test: The principles of assessment. In T. Swanwick (Ed.), *Understanding Medical Education. Evidence, Theory and Practice* (Second ed., pp. 243-254). USA: John Wiley and Sons, Ltd. .

Shumway, J M, & Harden, R M. (2003). AMEE Guide No. 25: The Assessment Of Learning Outcomes For The Competent And Reflective Physician. *Medical Teacher, 25*(6), 569-584. doi: doi:10.1080/0142159032000151907

Sibbald, D, & Regehr, G. (2003). Impact On The Psychometric Properties Of A Pharmacy OSCE: Using 1st-Year Students As Standardized Patients. *Teaching And Learning In Medicine, 15*(3), 180-185. doi: 10.1207/S15328015TLM1503_06

Singh, R, Hickner, J, Mold, J, & Singh, G. (2014). "Chance Favors Only The Prepared Mind": Preparing Minds To Systematically Reduce Hazards In The Testing Process In Primary Care. *JOURNAL OF PAtient Safety, 10*(1), 20-28.

Smee, S. (2003). Skill Based Assessment. *British Medical Journal, 326*(7391), 703-706.

Smee, S, Dauphinee, W, Blackmore, D, Rothman, A, Reznick, R, & Des Marchais, J. (2003). A Sequenced OSCE for Licensure: Administrative Issues, Results And Myths. *Advances In Health Sciences Education, 8*, 223-236.

Smith, C D, Worsfold, K, Davies, L, Fisher, R, & McPhail, R. (2011). Assessment Literacy And Student Learning: The Case For Explicitly Developing Students 'Assessment Literacy'. *Assessment & Evaluation In Higher Education, 38*(1), 44-60. doi: 10.1080/02602938.2011.598636

Southgate, L, Campbell, M, Cox, J, Foulkes, J, Jolly, B, McCrorie, P, & Tombleson, P. (2001a). The General Medical Council's Performance Procedures: The Development And Implementation Of Tests Of Competence With Examples From General Practice. *Medical Education, 35*, 20-28. doi: 10.1046/j.1365-2923.2001.00003.x

Southgate, L, Cox, J, David, T J, Hatch, D, Howes, A, Johnson, N, Jolly, B, Macdonald, E, McAvoy, P, McCrorie, P, & Turner, J. (2001b). The Assessment Of Poorly Performing Doctors: The Development Of The Assessment Programmes For The General Medical Council's Performance Procedures. *Medical Education, 35(Suppl. 1)*(2-8).

Stiggins, R J. (1997). Dealing With The Practical Matter Of Quality Performance Assessment. *Measurement In Physical Education And Exercise Science, 1*(1), 5-17. doi: 10.1207/s15327841mpee0101_1

Sudan, R, Clark, P, & Henry, B. (2015). Cost And Logistics For Implementing The American College Of Surgeons Objective Structured Clinical Examination. *American Journal Of Surgery, 209*(1), 140-144. doi: 10.1016/j.amjsurg.2014.10.001

Sutnick, A I, Friedman, M, Stillman, P L, Norcini, J J, & Wilson, M P. (1994). International Use Of Standardized Patients. *Teaching And Learning In Medicine, 6*(1), 33-35. doi: 10.1080/10401339409539640

Swanson, D, Clauser, B E, & Case, S. (1999). Clinical Skills Assessment with Standardized Patients in High-Stakes Tests: A Framework for Thinking about Score Precision, Equating, and Security. *Advances in Health Sciences Education, 4*, 67-106.

Swanson, D B, & Norcini, J J. (1989). Factors Influencing Reproducibility Of Tests Using Standardized Patients. *Teaching and Learning in Medicine, 1*(3), 158-166. doi: 10.1080/10401338909539401

Tavakol, M, & Dennick, R. (2011). Post-Examination Analysis Of Objective Tests. *Medical Teacher, 33*(6), 447-458. doi: 10.3109/0142159X.2011.564682

Tavakol, M, & Dennick, R. (2012). Post-Examination Interpretation Of Objective Test Data: Monitoring And Improving The Quality Of High-Stakes Examinations: AMEE Guide No. 66. *Medical Teacher, 34*(3), e161-175. doi: 10.3109/0142159X.2012.651178

Tavakol, M, & Sandars, J. (2014). Quantitative And Qualitative Methods In Medical Education Research: AMEE Guide No 90: Part I. *Medical Teacher, 36*(9), 746-756. doi: 10.3109/0142159X.2014.915298

Taylor-Adams, S, Brodie, A, & Vincent, C. (2008). Safety Skills for Clinicians: An Essential Component of Patient Safety. *Journal Of Patient Safety, 4*(3), 141-147.

Taylor, F W. (1911). *The Principles of Scientific Management* (PreLinger Library ed.). New York and London: Harper & Brothers Publishers.

Taylor, J S, Hunter, N, Basaviah, P, & Mintz, M. (2012). Developing A National Collaborative Of Medical Educators Who Lead Clinical Skills Courses. *Teaching And Learning In Medicine, 24*(4), 361-364. doi: 10.1080/10401334.2012.730452

Ten Cate, O, Snell, L, & Carraccio, C. (2010). Medical Competence: The Interplay Between Individual Ability And The Health Care Environment. *Medical Teacher, 32*(8), 669-675. doi: 10.3109/0142159X.2010.500897

Thomas, J, Schultz, T, Hannaford, N, & Runciman, W. (2011). Mapping The Limits Of Safety Reporting Systems In Health Care - What Lessons Can We Actually Learn? *Medical Journal of Australia, 194*, 635-639.

Tigelaar, D E, Dolmans, D H, de Grave, W S, Wolfhagen, I H, & van der Vleuten, C P. (2006). Portfolio As A Tool To Stimulate Teachers' Reflections. *Medical Teacher, 28*(3), 277-282. doi: 10.1080/01421590600607013

Townsend, P D, Christensen, M G, Kreiter, C D, & zumBrunnen, J R. (2010). Investigating The Use Of Written And Performance-Based Testing To Summarize Competence On The Case Management Component Of The NBCE Part IV-National Practical Examination. *Teaching And Learning In Medicine, 22*(1), 16-21. doi: 10.1080/10401330903445737

Tsai, T C, & Harasym, P H. (2010). A Medical Ethical Reasoning Model And Its Contributions To Medical Education. *Medical Education, 44*(9), 864-873. doi: 10.1111/j.1365-2923.2010.03722.x

Turner, J L, & Dankoski, M E. (2008). Objective Structured Clinical Exams: A Critical Review. *Family Medicine, 40*(8), 574-578.

Tversky, A, & Kahneman, D. (1974). Judgement Under Uncertainty: Heuristics And Biases. *Science, New Series, 185*(4257 (Sep.27. 1974)), 1124-1131.

Valentino, J, Donnelly, M B, Sloan, D A, Schwartz, R W, & Haydon, R C I. (1998). The Reliability Of Six Faculty Members In Identifying Important OSCE Items. *Academic Medicine, 73*(2), 204-205.

Vallevand, A L C. (2008). Reliability, Validity And Sources Of Errors In Assessing Physician Performance In An Objective Structured Clinical Examination: A Generalizability Theory Analysis. (Doctor of Philosophy), University of Calgary, Calgary, Alberta.

van der Vleuten, C P M. (1996). The Assessment of Professional Competence: Developments, Research and Practical Implications. *Advances in Health Sciences Education, 1*, 41-67.

van der Vleuten, C P M. (2000). Validity Of Final Examinations In Undergraduate Medical Training. *British Medical Journal, 321*(7270), 1217-1219. doi: 10.1136/bmj.321.7270.1217

van der Vleuten, C P M, Norman, G R, & De Graaff, E. (1991). Pitfalls In The Pursuit Of Objectivity: Issues Of Reliability. *Medical Education, 25*, 110-118.

van der Vleuten, C P M, & Schuwirth, L W T. (2005). Assessing Professional Competence: From Methods To Programmes. *Medical Education, 39*(3), 309-317. doi: 10.1111/j.1365-2929.2005.02094.x

van der Vleuten, C P M, Schuwirth, L W T, Scheele, F, Driessen, E W, & Hodges, B. (2010). The Assessment Of Professional Competence: Building Blocks For Theory Development. *Best Practice And Research. Clinical Obstetrics And Gynaecology, 24*(6), 703-719. doi: 10.1016/j.bpobgyn.2010.04.001

van der Vleuten, C P M, & Swanson, D B. (1990). Assessment Of Clinical Skills With Standardized Patients: State Of The Art. *Teaching And Learning In Medicine, 2*(2), 58-76. doi: 10.1080/10401339009539432

van Zanten, M, Boulet, J R, & McKinley, D W. (2003). Correlates Of Performance Of The ECFMG Clinical Skills Assessment: Influences Of Candidate Characteristics On Performance. *Academic Medicine, 78*(10/ October Supplement), S72-74.

Vargas, A L, Boulet, J R, Errichetti, A, van Zanten, M, Lopez, M J, & Reta, A M. (2007). Developing Performance-Based Medical School Assessment Programs In Resource-Limited Environments. *Medical Teacher, 29*(2-3), 192-198. doi: 10.1080/01421590701316514

Varkey, P, Natt, N, Lesnick, T, Downing, S, & Yudkowsky, R. (2008). Validity Evidence For An OSCE To Assess Competency In Systems-Based Practice And Practice-Based Learning And Improvement: A Preliminary Investigation. *Academic Medicine, 83*, 775-780.

Wakeford, R, Southgate, L, & Wass, V. (1995). Improving Oral Examinations: Selecting, Training, And Monitoring Examiners For the MRCGP. *British Medical Journal, 311*(7010 (Oct.7, 1995(), 931-935.

Walsh, K, Levin, H, Jaye, P, & Gazzard, J. (2013). Cost Analyses Approaches In Medical Education: There Are No Simple Solutions. *Medical Education, 47*, 962-968. doi: 10.1111/medu.12214

Walters, K, Osborn, D, & Raven, P. (2005). The Development, Validity And Reliability Of A Multimodality Objective Structured Clinical Examination In Psychiatry. *Medical Education, 39*(3), 292-298. doi: 10.1111/j.1365-2929.2005.02091.x

Wamsley, M A, Julian, K A, O'Sulivan, P, Satterfield, J M, Satre, D D, McCance-Katz, E, & Batki, S L. (2013). Designing Standardized Patient Assessments To Measure SBIRT Skills For Residents: A Literature Review And Case Study. *Journal Of Alcohol & Drug Education, 57*, 46-65.

Wang, F, & Hannafin, M J. (2005). Design-Based Research And Technology-Enhanced Learning Environments. *Educational Technology, Research And Development, 53*(4), 5-23.

Whitehead, C R, Kuper, A, Hodges, B, & Ellaway, R. (2015). Conceptual And Practical Challenges In The Assessment Of Physician Competencies. *Medical Teacher, 37*(3), 245-251. doi: 10.3109/0142159X.2014.993599

Wilkinson, T, Frampton, C, Thompson-Fawcett, M, & Egan, T. (2003). Objectivity in Objective Structured Clinical Examinations: Checklists Are No Substitute for Examiner Commitment. *Academic Medicine, 78*(2), 219-223.

Wong, B M, Levinson, W, & Shojania, K G. (2012). Quality Improvement In Medical Education: Current State And Future Directions. *Medical Education, 46*(1), 107-119. doi: 10.1111/j.1365-2923.2011.04154.x

Yeates, P, O'Neill, P, Mann, K, & Eva, K W. (2013). 'You're Certainly Relatively Competent': Assessor Bias Due To Recent Experiences. *Medical Education, 47*(9), 910-922. doi: 10.1111/medu.12254

Yudkowsky, R, Park, Y S, Riddle, J, Palladino, C, & Bordage, G. (2014). Clinically Discriminating Checklists Versus Thoroughness Checklists: Improving The Validity Of Performance Test Scores. *Academic Medicine, 89*(7), 1057-1062. doi: 10.1097/ACM.0000000000000235

Zabar, S, Kachur, E, Kalet, A, & Hanley, K. (2013). Objective Structured Clinical Examinations. Ten Steps To Planning And Implementing Osces And Other Standardized Patient Exercises (First ed.). New York, USA: Springer Science+Business Media New York.