Selection into Vocational Medical Training

by

Scott Sypek

Thesis
Submitted to Flinders University
for the degree of

Master of Clinical Education
College of Medicine and Public Health
24th January 2019

CONTENTS

SUMMARY

Selection into vocational medical training is high stakes and competitive. A process is needed to select successful trainees from a large pool of applicants. Through a literature review and mixed methods analysis of a local case study, this study sought to identify the factors that make an effective selection process for vocational medical training.

Ultimately an effective process will select trainees that are successful in training and in becoming competent specialists. However, in this context, there are few meaningful measures of trainee performance, most trainees will eventually complete training and there is a low attrition rate which makes predictive validity studies difficult. Instead, other indices are used as proxies for the effectiveness and quality of selection tools and processes. These include reliability, various types of validity, acceptability and feasibility. No one tool is able to perform well across all these areas. In fact, beyond the type of tool, there are several factors that determine a tool's utility in selection. These include the constructs measured by the tool and how these relate to the purpose of selection, the content, the format, the scoring system applied, and the number and training of assessors. Typically studies in this area are case reports of selection processes that focus on optimising the psychometric properties of selection tools. A gap in the literature is that there is no accepted standard theoretical or conceptual framework to guide selection process design and implementation.

The methods used to combine tool data and make selection decisions contribute to determining whether a selection process is effective. Many case reports in the literature and the local case study use a reductionist approach to decision-making. Information collected from the different selection tools is converted to a numerical score, which is summed to develop a rank list of applicants. This approach allows applicants to compensate for poor performance in one tool with better performance in another and therefore the weighting applied to each selection tool score has a significant influence on selection outcome. The quantification of qualitative data collected means that in this reductionist algorithm, valuable information about each applicant is lost when making final selection decisions. Constructing a selection process that considers all of these factors is complex and frameworks for designing and implementing selection processes are needed.

Assessment in medical education has faced many of the same challenges seen in selection. A Programmatic Assessment framework has been proposed to aid assessment practices. This framework involves the multi-method systematic collection of data about a learner, the careful selection of tools mapped to curriculum outcomes, and procedures for collating information collected about learners. The local case study is viewed through the lens of Programmatic Assessment. Utility of Programmatic Assessment principles to design a selection process provides a means to map domains to selection tools, combine information gathered on each applicant and facilitate decision making processes.

DECLARATION

I certify that this thesis does not incorporate without acknowledgment 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.

Scott Sypek

Date........16/9/2018...........

ACKNOWLEDGEMENTS

I would like to acknowledge the support, tolerance and professionalism of my two supervisors Julie Ash and Ruth Sladek. They have great expertise in different areas and were able to guide me through this research journey knowledgably, patiently and with agility. When road blocks seemed insurmountable we were able to overcome them with hard work and some hard discussions. I have learnt a lot from you both.

Thank you to my colleagues for support through this project Natalie Michael, Kylee Nuss, Stacey McPherson, David Everett and David Baulderstone.

To Megan, the most patient, supportive and talented person I know. Thank you for love, comfort, space, motivation, encouragement and tolerance. And to Audrey and Sebastian – for love and laughs, singing, movie nights, Blockus, Jenga, chess, charades, scootering… and for reorientation and grounding.

This study was undertaken as part of a Masters Thesis supported by the Australian Government Research Training Program Scholarship.

FIGURES AND TABLES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

1. INTRODUCTION

Australians enjoy high quality healthcare that relies upon on a highly skilled and trained healthcare workforce. Medical specialists are a key component of this workforce, who provide care to patients in both hospital and community settings. To become a medical specialist, doctors must apply and be selected into a vocational training program through a medical specialty college. This thesis examines selection into vocational medical training.

The Medical Training Landscape

Qualifying as a medical specialist requires significant commitment and many years of training. Figure 1 provides a snapshot of the medical training landscape in Australia. Australia has 18 University based medical schools, currently training around 16,000 medical students. Each medical school has its own selection process. Graduates must complete a compulsory one-year internship in order to qualify for general registration as a medical practitioner. Following their internship, most junior doctors spend a further one or two years working in clinical settings before applying for a vocational training position. Some doctors do not enter vocational training and continue working in hospital and community settings in supervised non-vocational positions.

Governance of vocational medical training rests with the medical specialty colleges. Applicants to vocational training programs may be interns, junior doctors working in prevocational training positions, or doctors working in non-vocational roles. There are also pathways for overseas trained doctors to enter training. Training times vary between programs, usually from three to seven years. Trainees who complete all training requirements are awarded a Fellowship from their medical specialty college. They can then obtain specialist registration with the Australian Health Practitioner Regulation Agency (AHPRA) and can work independently without supervision.

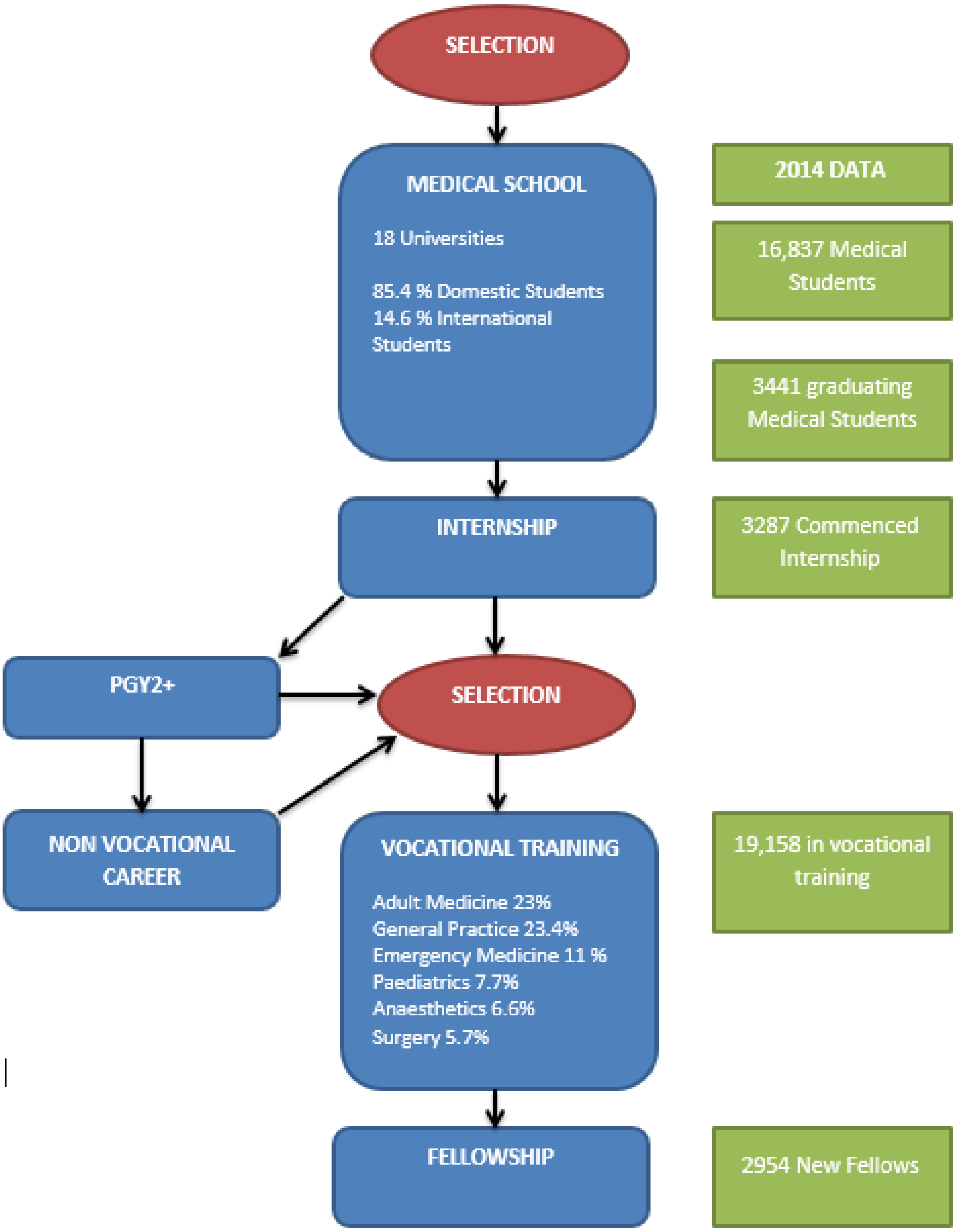


Figure 1: Medical training in Australia
PGY – Post Graduate Year
Note: Based on information in the Medical Training Review Panel Report (2015)

The number of graduating medical students in Australia has increased rapidly over the last decade. Graduate numbers have increased from around 1400 in 2000 to over 3600 in 2015 (Medical Deans Australia and New Zealand, 2015). Most of these graduates aim to pursue vocational medical training which has resulted in increased competition for training places.

There are a limited number of vocational training positions available. The number of training positions is limited by the capacity of health services to employ trainee doctors and by the colleges' ability to supervise and train them. In 2014 there were 19,158 doctors across all stages of speciality training (Medical Training Review Panel, 2015). For several reasons, it is difficult to calculate how many specialty training positions are available to junior doctors in Australia seeking to enter a programme. Some colleges do not limit the number of training positions but accredit training posts in health services that may or may not be filled. Data on the number of trainees include doctors who may be part of more than one training program. Additionally, some training programs only open up new positions when current trainees have obtained their fellowship. Nevertheless, with the demand for training places and the need for specialists, the number of medical graduates and availability of vocational training places is incongruous. In 2014, there were 3,385 doctors in their first year of vocational training. The same year there were 3,549 medical graduates (Medical Training Review Panel, 2015). With the persistent growth in the number of medical school graduates, competition for vocational training positions is likely to continue to increase.

Stakeholders in vocational medical selection are a broad group. The applicants need to know that selection processes are fair and transparent and that there are opportunities for a wide range of applicants to be successful (Kelly, Patterson, O'Flynn, Mulligan, & Murphy, 2018). The medical specialty colleges invest significant time and money in training. They will want to select people likely to complete training and avoid selecting those who will have difficulty. There is also a need for selection processes to be robust and defensible against legal challenge. Employers (hospitals, general practices and public health departments) require a stable and competent workforce to deliver healthcare to the public. The hospitals, along with selection coordinators require processes that are feasible in terms of the resources they have available to conduct them. The community (patients, community health providers, general public) need to know those selected into medical speciality training are of an appropriate standard to provide health care now and in the future.

Selection Processes

A selection process is needed to allow doctors into training who will succeed in completing the program and exclude those who are likely to have difficulty (Roberts & Togno, 2011). It is important not to select the wrong person as this can have negative impacts for the training hospitals, the individual and most importantly the patients (Patterson, Ferguson, & Knight, 2013). Selection into vocational training is a high stakes process for individual applicants. Once selected there are generally low attrition rates and a career as a medical specialist often means achieving a high level of income and social status (Prideaux et al., 2011).

Each medical speciality college has a different process in place to select trainees into their programs and these were subject to review twenty years ago (Brennan, 1998). The Medical Training Review Panel (MTRP) was established in 1996 by the Commonwealth Government to oversee the implementation of *The Health Insurance Amendment Act (No. 2, 1996).* The role of this body included reviewing how medical specialty colleges select their trainees. The panel engaged Dr Peter Brennan to investigate and report on the medical specialty training landscape in Australia. The review was prompted by several concerns about the selection into vocational medical training that included:

- Colleges intentionally limiting intake of trainees to protect the financial and market position of existing fellows
- Inconsistencies between college selection policy and actual practices
- Confusion around the role of the employing hospital and college in selecting trainees
- For some colleges there was no attempt to limit numbers in training
- Perceptions about lack of transparency for the selection process

The Panel surveyed medical specialty colleges across Australia and described the important issues around medical training selection at the time. Thus, the so called *Brennan Report* entitled *Trainee Selection in Australian Medical Colleges* was published in 1998. The report recommended a "Best Practice Framework for Trainee Selection" that has become known as the Brennan Principles and has been adopted by some of the colleges. These Principles are summarised below and include the following:

- There should be a clear statement of principles underpinning selection to select the best possible applicants and produce the best possible practitioners
- The process should be legal and accountable
- Eligibility criteria should be clearly stated
- Limits to the number of training positions should be declared openly
- Applications should be written in a standardised proforma
- Referees' reports should be standardised
- The selection committee should be an appropriate size, have training, be held accountable and have the confidence of the applicant, the profession and community
- The selection process should be valid, reliable, feasible and evaluation should be built into the process
- Selection criteria should be documented, published and be objective and quantifiable
- Selection committees should rank applicants
- The process should be documented
- All applicants should be given feedback
- There should be an appeals process

Many of these principles remain to be fully realised. Faced with a large number of applicants for a limited number of positions, specialty colleges need to design and implement a selection process that attempts to meet the Brennan Principles.

Systems for selection into vocational training vary by college and have evolved over time. The traditional approach of submitting a letter, curriculum vitae (CV), followed by a face to face interview and a check of references has been used across the world (Goodwin et al., 2014; Prideaux et al., 2011). There are now many more tools available. However, there are few studies evaluating the predictive validity of selection tools in vocational selection (Gale et al., 2010; Patterson, Rowett, et al., 2016) that could guide tool selection. Coordinators of selection must decide the relevant knowledge, skills and abilities required for the role (Patterson, Ferguson, et al., 2013) and what tools should be used and how. Panel interviews and referee reports are popular among applicants but have been shown to lack reliability and validity (Patterson, Knight, et al., 2016). Multiple mini interviews (MMIs) have better reliability but the construct validity is uncertain, and they require greater resources to administer (Dore et al., 2010).

Selection coordinators must also decide on a decision-making process to use the information from tools in making selection decisions. Although it is not always clear in published literature how this is done, the typical approach is to apply a score for each tool used and add the scores together to create a rank list (Bandiera & Regehr, 2004; Goodwin et al., 2014; Shulruf et al., 2018). The validity of summing scores from highly variable tools that measure different constructs could be questioned, however this approach is widely used to create a rank list of applicants.

There are many factors to consider when designing a selection process and there appears to be no accepted best practice model. What applicant qualities should be selected for? What is the relative importance of assessed domains such as academic versus interpersonal skills? Who should decide this? Should a short-listing process be used? What proportion of applicants should be short listed? Which selection tools should be used? How many different methods are optimal? How should different tools be weighted? Which methods are best at predicting future performance? Is there a role for using personal references? Should interviews be used? Who should be the interviewers? What training do they require? Choices made in answering each of these questions will influence the effectiveness of a selection process. The lack of clear theoretical and conceptual frameworks to guide these choices has been identified as a gap in the current selection literature (Roberts et al., 2017).

A number of frameworks have been proposed but there is no established standard to guide selection process design. The Brennan Principles provide a set of aspirational goals but without the instruction to achieve them (Brennan, 1998). It has been suggested that selection systems be modelled on educational assessment practices (Prideaux et al., 2011) and there are examples of selection processes using the principles of competency based medical education (Gale et al., 2010; Patterson, Tavabie, et al., 2013). These processes use principles of organisational psychology; desirable attributes for trainees are identified through a range of methods including job analysis and interviews with stakeholders. Selection tools are then chosen and mapped to measure the attributes (Randall, Davies, Patterson, & Farrell, 2006; Vermeulen et al., 2014). However, studies using a competency based education model are limited and predictive validity studies are not available to judge whether the extra resources required for this process result in the selection of better trainees. The precise elements that make a selection process effective have not yet been clearly defined.

This thesis set out to answer the research question:

What is an effective process for selecting applicants into vocational medical training programs?

Several elements in this question require definition. Predictive validity is a key factor in determining the *effectiveness* of a selection process. Other considerations include fairness, reliability, the acceptability to stakeholders and feasibility. Selection should be considered beyond simply the tools that are used. A *selection process* encompasses how criteria are established, the scoring systems used, the weighting of tools, the staff involved and their training, and the evaluation and quality assurance processes.

This thesis aims to explore this question through a literature review and mixed methods analysis of a case study of selection in vocational medical training. Through examining the evidence from the literature and the issues faced in a local case study, the intention is to establish a framework for vocational selection that is effective.

2. LITERATURE REVIEW

The aim of the literature review was to explore the current practice in vocational medical selection with a view to unearthing the features of an effective process. First the search strategy undertaken is outlined. Next, a summary of the key selection methods used is presented. A sample of case reports typical of those appearing in the literature is described to illustrate the common issues in vocational medical selection. Finally, the gap in the current literature around lack of theoretical frameworks to underpin selection is presented.

There is a wide literature examining the processes and outcomes of medical *student* selection. While there are lessons to be learnt from this literature there are important differences in selection for *vocational medical training.* Applicants to vocational medical programs are doctors who have completed medical school and achieved the necessary requirements for registration as a medical practitioner. There is a set of skills and attributes that are assumed to have been obtained through this process. Medical school applicants are also a more heterogeneous group and less can be presumed about their past experience and ability at the commencement of the selection process. However, both processes are high stakes and require a process to move from a large number of applicants to a much smaller number of successful trainees. Therefore, it was important to construct a search strategy that would find published articles about vocational medical training selection, as well as relevant key articles around medical student selection.

It was not the intention to undertake a systematic review of the selection literature, rather an iterative process underlined the construction of this literature review. The initial literature search used Medline, PubMed, Scopus, Proquest and Informit databases using combinations of the search terms in Figure 1. This identified 1433 articles which, after title and abstract review, was reduced to a list of 235 papers considered relevant. All 235 papers were retrieved and each manuscript was reviewed. Those papers that contributed to exploring the effectiveness of vocational medical selection are cited in this literature review. Many papers were local case studies in an irrelevant context and offered little over the studies that are included. A weekly alert was set up for each database to identify any new articles that met search criteria. Over the course of the project, using updated database searches, email alerts and reviews of reference lists, further articles were identified. A significant proportion of the vocational medical selection literature is concerned with the measurement tools (also referred to as *methods* or *instruments*) available. Often these are presented in case reports of local selection processes, with emphasis on the psychometric properties of the tools and commentary on the logistics of the selection process. This next section provides a description of the tools commonly used and summarises evidence for their use. Six case reports that illustrate the major issues in vocational medical selection are then described.

applicant, candidate, aspirant, interview, multiple-mini-interview, MMI, reference, referee, rank, shortlist, domain, situational judgement test, SJT, admission, select, preselect, entrance, entry, recruit, specialty, subspecialty, fellowship, graduate, postgraduate, surgical, paediatric, pediatric, anaesthetic, radiology, rheumatology, haematology, cardiology, dermatology, general practice, family practice, training

Figure 1: Search terms used in initial database search

Selection Instruments

For those charged with coordinating a selection process there are a number of selection tools available to choose from. The terminology used to discuss the psychometric properties of selection tools is diverse and inconsistent. There is debate about the validity of particular selection tools with varied interpretation of what this means. Validity refers to the degree to which the conclusions reached from the results of a test are meaningful for a specific purpose (Cook & Beckman, 2006). Table 1 provides an overview of the types of validity referred to in the literature. The term, *validity*, is often not defined in the selection literature but is most commonly used to describe construct validity, or a combination of a number of validity types. In this literature review when *validity* is used without a defining prefix, it is used broadly, to refer to the meaningfulness and value of the information gathered from a tool, with respect to its purpose in selection.

Table 1

Validity in selection

Face Validity	The appearance that a selection tool is valid without subjecting it to any practical testing
Criterion Validity	The extent to which the results from a selection tool are related to an outcome
Concurrent Validity	The extent to which the results from a selection tool are related to an outcome measured at the same time (e.g. another tool in the selection process)
Predictive Validity	The extent to which the results from a selection tool are related to an outcome measured in the future (e.g. future performance in training)
Construct Validity	The degree to which a tool measures the construct it is intending to measure
Content Validity	Refers to the relevance of the selection tool content to the target role
Political Validity	The extent to which stakeholders view that a selection tool is appropriate and acceptable
Social Validity	A measure of the social impact of selection as it relates to perceived and actual fairness

(Burgess, Roberts, Clark, & Mossman, 2014; Cook & Beckman, 2006; Kelly et al., 2018; Messick, 1989; Patterson, Ferguson, et al., 2013)

Two recent systematic reviews have presented the evidence for the most commonly used selection tools in both medical student (Patterson, Knight, et al., 2016) and vocational medical selection (Roberts et al., 2017), from here on referred to as the ‘Patterson review’ and the ‘Roberts review.’ Both papers are recent, comprehensive reviews of the extant literature, and subsequently are referenced often throughout this chapter. Therefore it is important to provide an overview of these important reviews.

The Patterson review was undertaken to summarise the research evidence for selection methods used to select into medical school training and excluded articles on vocational medical training. A total of 194 articles met inclusion criteria and selection tools were considered under four research questions concerning: effectiveness, procedural issues, acceptability and cost effectiveness. The review is relevant for this thesis as it summarises the literature on a number of the same tools used in vocational medical selection. Importantly, the review considers a range of other factors beyond predictive validity that are pertinent to judging whether a selection tool is effective. These include logistical issues of using the tool, efficiency in terms of cost and resources and stakeholder views. The process of assessing the quality of the included studies was not described except to say that in general study quality was low, dominated by cross sectional studies and further reporting on quality was ‘beyond the scope’ of the review (p40). The review has been criticised due to concerns that a significant volume of peer reviewed published literature on SJTs has been produced by the authors of the review who are members of the Work Psychology Group and stand to potentially benefit financially from the advancement of the use of SJTs (Adam et al., 2017). Whilst it is worthwhile bearing this issue in mind, nevertheless, a number of papers not authored by affiliates of the Work Psychology Group support the use of SJTs (Lievens, Buyse, & Sackett, 2005; Roberts et al., 2014).

The Roberts review deals specifically with vocational medical selection and considers selection beyond the evidence for individual tools. The review identified 116 articles on vocational medical training selection with 89 considered to be high quality based on the strength of conclusions drawn and the global impression of the reviewers. Fifty papers were concerned with selection tools and van der Vleuten's utility index (van der Vleuten, 1996) was used to evaluate tool effectiveness. This index is used in medical education assessment and considers the utility of an assessment method in terms of a combination of its reliability, validity, educational impact, acceptability, feasibility and cost effectiveness. These are remarkably similar criteria to those used in the Patterson review. In addition to considering the evidence for selection tools, the Roberts review also considers the frameworks in which selection tools operate. They identify two types of framework in practice: the first is the selection system typically used in the United States that is more subjective and based on locally defined criteria and gives high regard to previous academic results. This is distinguished from a selection system that is competency based and uses multiple methods of selection. Selection tools used in both types of frameworks are considered in this literature review. Table 2 summarises the findings of the Patterson and Roberts reviews, and shows the broad tools used have variable levels of psychometric rigour.

Table 2

Summary of findings for Roberts and Patterson reviews

Tool	Roberts et al (2017) Vocational Medical Selection	Patterson et al (2016) Medical Student Selection
Multiple Mini Interview	Relatively high reliability for an observed assessment. Good predictive validity but data supporting validity often context-specific. Concerns about cost.	Improved reliability over single interview Concerns about construct validity Relatively expensive to design and implement
Structured Interview	Good reliability – improved by interviewer training and standardised scoring systems. Limited generalisability to other settings.	Reliability and validity improved with standardised questions, trained interviewers and appropriate scoring system. Favourable amongst applicants. Resource intensive
Situational Judgement Tests	Favourable reliability and predictive validity. Results yet to be reproduced in other settings. Concerns about expensive development costs.	Improved validity over IQ and personality tests, Costly to design. Useful in high volume selection as can be delivered online or machine marked.
Personality testing	Little justification in developing personality testing based on current frameworks.	High risk for susceptibility to faking or coaching Can be used in association with interviews rather than as a standalone instrument.
Letters of recommendation and references	Limited evidence on reliability and validity	Use of references is widespread Little research supporting validity or Reliability
Personal statement/ CV	No predictive validity between the CV and subsequent performance.	Personal statements are susceptible to coaching. High acceptability amongst applicants.
Selection Centres (SC)	Concerns over cost effectiveness Predictive validity unclear Mixed results for acceptability	Expensive to design and implement. Sparse evidence of the predictive validity of SCs

The following section uses the two systematic reviews and the wider literature to describe the main tools used in vocational medical selection and summarises the evidence for their use: interviews, situational judgment tests, personality testing, referee reports, curriculum vitae and selection centres.

Interviews

The interview is a commonly used method in vocational medical selection. Interviews involve one or more interviewers asking questions to an applicant and assessing their responses. Interviews are classified by their structure (unstructured or structured) and can vary in their duration, content, number of interviewers and mode of administration (Patterson, Ferguson, et al., 2013). Multiple Mini-Interviews (MMIs) are a specific format of structured interview that usually requires the applicant to progress through a number of interview stations of 5-10 minutes duration, each with a different interviewer or set of interviewers. Rather than one interaction with a single interview panel, an applicant has several independent interactions with different interviewers.

The literature contains much discussion about the preferred number of interviewers, stations and their duration to achieve optimal effectiveness, in both vocational and medical student selection (Dore et al., 2010; Eva et al., 2009; Yoshimura et al., 2015). Although the ideal number of MMI stations is context dependent, in general, increasing the number of stations, using rating scales and providing training for assessors are positive factors for enhancing reliability (Knorr & Hissbach, 2014). Interviews have high levels of acceptability among stakeholders (Patterson, Lievens, Kerrin, Munro, & Irish, 2013; Razack et al., 2009). Applicants have viewed MMIs as fair and allowing them the opportunity to demonstrate their strengths (Dore et al., 2010). In some studies applicants expressed a preference for interviews with the one panel (rather than MMIs) as it allowed greater opportunity for a personal connection (Razack et al., 2009; Soares et al., 2015). Both the Roberts and Patterson reviews report that MMIs probably offer improved reliability and validity over panel interviews. Where panel interviews are used, standardised questions, scoring and training of assessors improves reliability and validity.

Findings regarding the predictive validity and construct validity of interviews are mixed. There is some evidence that MMIs can predict performance in *end* of training assessments in general practice training (Roberts et al., 2014) although prediction for work-based assessments of trainee performance and in training exams, reports are contradictory. Some studies report positive correlations between interviews and in training performance (Olawaiye, Yeh, & Withiam-Leitch, 2006; Oldfield, 2013) and others negative or equivocal associations (Adusumilli et al., 2000; Bell, Kanellitsas, & Shaffer, 2002). The ability of MMIs to predict performance in licensure examinations (in the United States and Canada) is variable and there are only weak correlations between MMI scores and academic performance in medical school (Eva et al., 2009; Hofmeister, Lockyer, & Crutcher, 2009). MMIs are a *method* to collect information about an applicant and the internal processes within the interview will impact on the validity of that information. For example, when interviewers have access to the past academic results for applicants, there is a correlation between these results and interview scores, which is not observed when interviewers were blinded, indicating bias. (Swanson et al., 2005). The type and range of scoring scale used can be altered but how this affects validity is unclear (Knorr & Hissbach, 2014). Overall, the optimal factors to enhance the validity of interviews are not well defined.

Whilst interviews are generally a well-accepted selection method, their effectiveness as defined by predictive and construct validity is variable. The content, structure and scoring system of the interview will impact on validity (Eva, Macala, & Fleming, 2018) and therefore also on its relative effectiveness as a selection tool. Many questions remain about the factors required for optimal utilisation of interviews.

Situational Judgement Tests

Situational judgement tests (SJTs) are a measurement method where applicants are asked to consider hypothetical scenarios and select one or more responses from a suggested list of alternatives (Clevenger, Pereira, Wiechmann, Schmitt, & Harvey, 2001). Clinical and non-clinical scenarios can be used and response options can include knowledge based outcomes (*What should you do?)* and behavioural tendencies *(What would you do?)* (Patterson, Zibarras, & Ashworth, 2016). The scenario content, response format, test length and scoring system should reflect the purpose of selection process and the constructs being measured (Patterson, Rowett, et al., 2016). As assessments, SJTs have been used to measure non-academic domains such as empathy and integrity (Koczwara et al., 2012) and may be less susceptible to coaching (Patterson, Ashworth, et al., 2012).

When creating SJTs there are many variables that influence their reliability and validity. While the design and construction of SJTs can be a complicated, iterative process requiring considerable resources (Roberts et al., 2017), once designed, they can be delivered online, to large cohorts and machine marked, therefore potentially improving the efficiency of a selection process (Patterson, Knight, et al., 2016). Both the Patterson and Roberts reviews report favourable reliability and validity for SJTs in both medical student and vocational selection. This needs to be considered in the context of the definition of a SJT. Like the MMI, there are many variables in the SJT: the stem format (video/written text, visual), question design (behavioural/situational), response structure (single response/rank), and scoring rubric (Koczwara et al., 2012; Lievens et al., 2005; Patterson, Knight, et al., 2016). Each of these factors will impact on the reliability and construct validity of the SJT – and therefore its contribution to the effectiveness of the overall selection process.

Personality Tests

Personality tests have been used in selection for medical school but less commonly in vocational medical selection. These assessments usually consist of self-report questionnaires where applicants respond to a variety of statements or descriptors. Results of personality tests provide a report of an applicant’s personality traits that is generally based around five factors: extraversion, emotional stability, agreeableness, conscientiousness, openness to experience (Patterson, Ferguson, et al., 2013). There is no clear agreement about which personality traits should be selected for (Roberts et al., 2017). Conscientiousness has been reported to be a positive predictor of preclinical knowledge but a negative predictor of clinical skills (Ferguson, James, O'Hehir, Sanders, & McManus, 2003). A number of studies have found no correlation between particular personality traits and in training performance (Dawkins, Ekstrom, Maltbie, & Golden, 2005; Selber et al., 2014). Personality type has been shown to influence selection decisions with interviewers providing more favourable ratings to applicants of similar personality type (using Myers-Briggs-Type Indicator) (Quintero, Segal, King, & Black, 2009). Based on the literature review and the findings in the Roberts and Patterson reviews, there is no clear role for personality tests in vocational medical selection.

Referee Reports

Despite substantial evidence indicating that referee reports are not good predictors of later job performance, they are widely used in vocational medical selection processes (Ferguson et al., 2003; Stedman, Hatch, & Schoenfeld, 2009). A referee report usually requires an applicant to approach a previous employer or supervisor to complete a templated referee report and make written comments about the applicant. Letters of Recommendation and Dean’s letters are comparable tools used in selection in the United States (Prager, Myer, & Pensak, 2010). The discriminatory utility of referee reports is limited as referees tend to apply the same scores to applicants irrespective of ability. In a review of 241 letters of recommendation written for 78 applicants to an internship program in Texas, Stedman et al (2009) found that letters of recommendation are positively biased, with positive and negative comments distributed uniformly across all applicants. Less experienced referees tend to provide higher ratings to applicants (Beskind et al., 2014). While a standardised rather than narrative format is more feasible and acceptable to the writer, it can lead to inflated scores (Roberts et al., 2017). As most referee reports provide positive comments on applicants, those that provide negative information may be most useful for selection coordinators for selecting out (Patterson, Ferguson, et al., 2013). Despite the shortcomings of referee reports they are reported to be valued by both program directors and applicants (Makdisi, Takeuchi, Rodriguez, Rucinski, & Wise, 2011), suggesting they may be acceptable despite their lack of reliability and validity.

Curriculum Vitae

A curriculum vitae (CV) is widely used in recruitment across many industries. Although the content can vary, it is usually a self-reported account of an applicant's work history, educational qualifications, professional achievements and extra-curricular activities. Appraising the value of using CVs in selection is difficult as published studies do not often include details of the contents of the CVs or how they are used. The CV can be used purely as a shortlisting tool to determine who will be interviewed (Bandiera & Regehr, 2003) or discussed at interview with the applicant to explore its content further (Swanson et al., 2005). Some processes, such as general surgical trainee selection in Australia, require applicants to submit a highly structured CV specifying professional, academic and personal achievements scored according to set criteria (Oldfield, 2013). This can lead to applicants with greater years of experience achieving higher scores which may not necessarily reflect aptitude for training. Other processes are less prescriptive and rely on a global impression from an assessor to determine how items listed in the CV may be mapped to desired applicant attributes (Bandiera & Regehr, 2003). Assessing self-reported achievement is problematic as there is usually no process to confirm the veracity of an applicant's CV. In a cohort of applicants to gynaecology oncology fellowships, 30% of publication claims and 71% of reported manuscript submissions were unable to be verified (Frumovitz et al., 2012).

In summary, CVs are widely used in vocational selection despite limited evidence available on their reliability or validity. The Roberts review includes only one study that used CVs in vocational medical selection which found a negative correlation between CV rating and subsequent performance in training (Oldfield, 2013). Based on this they conclude there is little value in their use in this setting. There is insufficient detail in the literature regarding the content of CVs, scoring systems and guidelines given to applicants to properly appraise their effectiveness in vocational medical selection.

Selection Centres

Some selection processes use a ‘Selection Centre,’ where an applicant is assessed by a range of exercises and tools by multiple assessors, usually on the same day (Patterson, Knight, et al., 2016; Roberts et al., 2017). Assessment stations can include written tasks, group exercises, simulations and clinical skills stations (Gale et al., 2010; Mitchison, 2009). This format allows the applicant to be assessed on multiple occasions in multiple contexts by multiple assessors, which is reported to result in greater reliability, validity and applicant satisfaction with the process (Patterson, Ferguson, et al., 2013). Both the Patterson and Roberts reviews report that evidence for the effectiveness of selection centres is insufficient. As the evidence for the individual activities within a selection centre is mixed, it is difficult to appraise the ‘evidence’ for selection centres as whole. Examples of selection centres used in vocational medical selection are discussed later in this chapter (Burgess et al., 2014; Randall et al., 2006).

Each tool described is a *method* of gathering information, not a measurement itself, and the reliability, construct validity and utility, and thus the effectiveness, is therefore influenced by many factors. These include training of assessors, the content of the tool and how it is mapped to the purpose of the selection process, scoring rubrics used, the number of tools, their duration, and whether assessors are blinded to other information about the applicant (Eva et al., 2018; Goodyear, Jyothish, Diwakar, & Wall, 2007; Patterson, Zibarras, et al., 2016). The validity will also be determined by how the information gained is interpreted and used (Colliver, Conlee, & Verhulst, 2012). Validity is a property of the conclusions reached, not the selection tool itself(Baker, Wallace, Cooke, Alpert, & Ackerly, 1987). The published literature on vocational medical selection consists of many case reports that describe how these tools are implemented in local selection processes. These case reports give an insight into the challenges faced by coordinators of selection and how selection processes are constructed and implemented. A sample of six case reports was chosen to illustrate the pertinent issues in vocational medical selection. These are summarised in Table 3 and discussed below.

Common themes from case reports

There are several common themes that emerge from these six illustrative case reports (Table 3). The themes concern the way selection tools are chosen and used, the difficulties in performing predictive validity studies and the emphasis on the psychometric properties of selection tools.

Table 3: *Summary of case reports*

Authors and Setting	Selection Methods and Weighting	Resources	Decision-making Process	Effectiveness Indices Reported
Bandiera and Regehr (2003) Emergency Medicine, Canada 40 applicants for two positions	Application Package (50%): • Curriculum vitae (17.55%) • Personal Letter (17.5%) • Academic Transcript (7.5%) • Three letters of reference (7.5%). Interview (50%)	Each application package: 3 assessors 4 teams of 2 interviewers. (30 min briefing)	Weighted average of scores from interview and application package	Inter-rater reliability
Gale et al (2010) Anaesthetics, UK 143 applicants for 37 positions	Structured interview (25%) Portfolio presentation (25%) Simulation (25%) Oral presentation (25%)	2 days of assessor training 7 days of planning workshops	Summed score from all stations Global score on safety and professionalism to allow for veto	Inter-rater reliability Correlation between tools Stakeholder acceptability Predictive validity
Goodwin et al (2014) Orthopaedics, UK 498 applicants	Three interview stations. Station 1: discussion of portfolio (50%) Station 2: knowledge: (25%) Station 3: Communication skills (25%)	210 interviewers Online training module	Sum of interview scores	Applicant acceptability
Goodyear et al (2007) Paediatrics, UK 224 applicants, 123 interviewed	Shortlisting based on application: • Publications and presentations • Audit • IT and communication skills • Personal statement MMI (3 stations)	16 shortlisting assessors 12 interviewers Training time not specified	Final selection decision-making process not specified	Inter-rater reliability
Randall et al (2006) Paediatrics, UK 27 applicants for 10 positions	Assessment Centre: • Structured panel interview (43%) • Group exercise (19%) • Written reflection (19%) Simulated Consultation (19%)	3 hours assessor training 8.5 hours per applicant at selection centres	Selection based on final scores and qualitative discussion between assessors.	Correlation between tools Applicant acceptability
Roberts et al (2014) General Practice (GP), Australia 1382 applicants for 1200 positions	SJT (50%) MMI (50%) (6 stations, single interviewer)	254 interviewers at 11 assessment centres Interviewer workshop (time not specified)	Sum of SJT and MMI	Correlation between tools Reliability

The identification and mapping of selection domains

Selection processes have evolved to become more structured and regulated than in the past. Part of this formalisation has been the move to identify desirable attributes in trainees and choose tools to target these attributes or domains, similar to approaches used in a competency based medical education model (Patterson, Ferguson, & Thomas, 2008). Desirable attributes are identified through a variety of means that include job analysis (Randall et al., 2006), review of documents from medical colleges (Bandiera & Regehr, 2003) critical incident reviews (Patterson et al., 2008) and using clinician reference groups (Goodyear et al., 2007). These can be thorough and intensive undertakings. As an example, in a selection process for paediatric trainees in the United Kingdom (Randall et al., 2006), methods used to define trainee attributes included, job analysis by an occupational psychologist observing paediatric consultants, critical incident reviews, and interviews with doctors, nurses and patients. A total of 164 behaviour descriptions were obtained and then grouped into 14 competency domains. In a similar process, an anaesthetic training program (Gale et al., 2010) used a multi-method job analysis published in a previous study (Patterson et al., 2008) to identify attributes. Then, an expert panel of anaesthetists, a human resources specialist and an occupational psychologist reviewed this information at a one-day workshop and identified competency domains for selection: achievements, communication, working under pressure, organisation and planning, situational awareness/decision-making and team working. Selection tools were chosen to assess applicants against these attributes and competency domains.

Once attributes are identified, the next step is to map them to selection tools, it is not always clear how this is done. In the paediatric program described above, attributes were mapped across a number of *different* tools (Randall et al., 2006). They mapped competency domains to a structured interview *and* three exercises at a selection centre: a simulated consultation with a concerned parent played by a trained medical actor, a group exercise to discuss and prioritise competing tasks faced by a paediatric trainee, and a reflective written exercise based on the group task. In other case reports, the ‘mapping’ process consisted of altering the content across the *same* tool. This occurred in an orthopaedics training program in the UK (Goodwin et al., 2014) where selection criteria were mapped across three interview stations. One station dealt with the applicant’s portfolio, one with two components of knowledge (clinical and anatomical) and the third station with two components of communication skills. Each component was assessed against two domains (e.g. judgement under pressure, problem solving). In both these examples, the justification for choosing a tool for a particular attribute is not articulated. Although comprehensive processes are undertaken to identify attributes, the emphasis in the literature is on the content, structure and psychometric properties of tools and less so on why particular tools are used for specific attributes.

The need for predictive validity studies and associated challenges

A consistent theme in the selection literature is the desire to know whether selection processes and tools are selecting trainees who perform well in training and into their careers. Most case reports acknowledge the lack of long term validation studies (Goodyear et al., 2007; Randall et al., 2006; Roberts et al., 2014). The predictive validity of tasks in a selection centre for anaesthetic trainees (Gale et al., 2010) was explored looking at the correlation between selection centre scores and a workplace based assessment (in-theatre assessment, ITA) and also an annual review of progress score which rates professional and clinical skills. The authors report ‘reasonable’ correlation between selection centre scores and subsequent work performance scores with Pearson’s correlations between r=0.33 and r=0.48. However, they do not report whether scores in particular stations in the selection centre correlate with work performance. This would be useful information when assessing the value of individual selection centre components. They acknowledge that they cannot follow up non-selected applicants in a comparable way and that there are limitations to workplace-based assessments that do not discriminate between trainees’ performance well.

Validation studies are difficult in vocational medical selection. Unlike medical school where there are multiple exams and assessments to compare the performance of students, in vocational medical training there are far fewer standardised assessments (Patterson, Ferguson, et al., 2013). It is difficult to assess overall work performance in a standardised way. Even in medical school selection there is no clear established framework for assessing the success of selection processes (Patterson, Knight, et al., 2016).

An attempt to test predictive validity of a selection process for public health trainees highlights some of the challenges (Pashayan et al., 2015). The selection process used two psychometric tests, a situational judgement test, followed by a selection centre that included panel interviews, group and written exercises. They used binary outcome measures of success in training; pass at first attempt of two examinations and satisfactory outcome of the annual performance review. Even with a large cohort (n=274), predictive validity was difficult to measure because of low rates of differentiation between trainees. The exam was passed by 90% of trainees at first attempt and 84% had satisfactory annual review outcomes recorded. This is a common phenomenon, with attrition rates being low once doctors have entered speciality training.

Performance measures on which to base predictive validity studies are limited. Roberts et al (2017) looked at 27 studies that reported on whether selection tools could predict later performance in vocational medical training. The outcomes used to reflect performance were examinations (in training and end of training) and work-based assessments through vocational medical training. Several studies reported United States Medical Licensing Examination scores predicted performance in both in training and end of training examinations. Findings related to other predictor variables (medical school grades, honours society status, research experience) were mixed. Obtaining meaningful data on which to base predictive validity studies is problematic. In work based assessments, supervisor reports are often based on global ratings of performance over a period of time, are frequently founded on third party observations, and reports are regularly written with reference to just a few samples of performance (van der Vleuten, 1996). Supervisors can be reluctant to rate trainees as unsatisfactory due to concerns about repercussions (Ende, 1983). With both examinations and work-based assessments, there is potential for a ceiling effect with most trainees performing well (Barrett et al., 2015). It is also difficult to follow up comparable measures in applicants who were unsuccessful in the selection process (Gale et al., 2010).

Reporting psychometric measures as evidence of quality - Reliability

Reliability dominates the selection literature. However, the psychometrics used to report reliability are often inconsistent which makes comparisons between tools and selection processes difficult (Roberts et al., 2017). Some studies report interrater reliability to evaluate the consistency between assessors using the same tool to assess the same applicant (Adams et al., 2009; Gale et al., 2010; Goodwin et al., 2014). While consistency between assessors is important, variability in the perspective of assessors can also be seen as valuable. Bandiera and Regehr (2003) give an example of the value of diverse opinions about an applicant. They describe an applicant who does volunteer work in a homeless shelter. One assessor may consider this positively with regard to vocational medical training because it 'encourages understanding of an underserviced population and develops communication skills' (p 598), while another assessor might be ambivalent because it is not medical. Both opinions are valuable but the inconsistency in marking will weaken measures of reliability.

Reliability in a selection process is also reported in terms of the consistency between scores across a number of selection measures. This is often presented as correlations between different items used in a selection process (Dore et al., 2010; Randall et al., 2006; Roberts et al., 2014). In their Selection Centre, Gale et al (2010) report on correlations between scores for individual selection stations and the applicants' final selection score. They conclude that positive correlations between each station score and the final score indicate the stations form a coherent evaluation of applicant performance. However, positive correlations would be expected as the scores for each station contribute to the final selection centre score. Reliability is concerned with the reproducibility of an assessment measure where what is being measured is a stable construct (de Vet, Terwee, Knol, & Bouter, 2006). Variation between scores for different tools could indicate that the stations are actually assessing different constructs (Roberts et al., 2014) or a construct that is not stable (Schuwirth & van der Vleuten, 2006). Consistent scoring across stations could be interpreted either that the applicant has similar performance in a number of attributes or that there is redundancy in assessing the same thing over a number of stations. Unless there is appropriate construct validity of the selection instruments, the reliability of the assessors' scores has little meaning, and they run the risk of being *reliably wrong* about what they are purporting to measure (Patterson & Ferguson, 2012).

Reporting psychometric measures as evidence of quality – Political Validity

The stakeholder acceptability of a selection process (political validity) is another proxy for quality reported in selection literature (Adams et al., 2009; Gale et al., 2010; Mitchison, 2009). In the sample case reports this information was collected through a survey of applicants after they have finished selection tasks but before they have received notice of the outcome (Gale et al., 2010; Goodwin et al., 2014; Randall et al., 2006). When applicants have negative perceptions of the selection process this can impact of their level of engagement with the process (Burgess et al., 2014) as well as increasing the potential for applicants to make legal challenges to selection decisions (Koczwara et al., 2012). While many studies report on the opinions of applicants and interviewers, there is little information available in the views of other stakeholders, the hospitals, nursing and allied health staff or most importantly, the patients (Kelly et al., 2018). There is also often community and media interest in selection processes due the status that a career in medicine holds in society (Patterson, Lievens, Kerrin, Zibarras, & Carette, 2012).

Applicants' perception of fairness (social validity) is affected by how they view the relevance of the task. Interviews and MMIs have received favourable ratings from applicants (Dore et al., 2010; Gale et al., 2010; Goodwin et al., 2014). Simulations have been rated higher than interviews in terms of fairness and ability to demonstrate skills (Gale et al., 2010). This has been described as being consistent with the theory of procedural justice, that is, applicants favour selection methods perceived to be related to the job (Kelly et al., 2018). The positive ratings for selection centres that use simulated patients and clinical tasks also concur with this theory (Randall et al., 2006). Other case reports found different results with clinical practice-related questions being viewed negatively. Applicants to GP training perceived clinically-based MMI questions as assessing readiness for practice as a GP rather than their ability to enter training and are therefore judged unfair (Burgess et al., 2014). In general though, irrespective of the selection method used, content related to the expected work while within the program positively influenced the perception of fairness by the applicants.

Significant resources are required for selection

Significant resources are invested in contemporary vocational medical selection processes which is a change from processes of the past. For example, previously orthopaedic surgeons would meet 10 minutes before the first interview to decide on the questions and processes to use whereas nowadays 210 interviewers are all required to complete an online training module prior to meeting applicants (Goodwin et al., 2014). Recruitment professionals, including organisational psychologists and human resources experts, coordinate selection processes once run by a small group of senior clinicians (Gale et al., 2010; Randall et al., 2006). Staff involved in the process often undergo training courses and modules from a few hours (Randall et al., 2006) to a number of days (Gale et al., 2010) (Table 2). Regional and nationwide processes require standardised training for large numbers of assessors (Goodwin et al., 2014; Roberts et al., 2014). The resources allocated to selection are an indication of the pressure of selection coordinators to have a credible process for high stakes selection. Involving recruitment professionals and standardised training can enhance reliability and face validity but must be balanced against costs so the process remains feasible (Koczwara et al., 2012).

Decision-making algorithms are unclear

Selection decisions are made based on information obtained from selection tools – how these decisions are reached varies and is not always well defined. Shortlisting processes may or may not be considered in the final selection decision and in many case reports it is unclear how the final outcome was determined (Goodyear et al., 2007; Shulruf et al., 2018). A reductionist approach is widely used. This is where applicants receive a score for each selection tool used, the weighted scores are summed and a rank list developed and used to offer training positions (Bandiera & Regehr, 2003; Goodwin et al., 2014; Roberts et al., 2014). In this approach it is possible to compensate for poor performance on one tool with good performance on another (Shulruf et al., 2018). The weighting given to each selection tool is an important consideration it will influence selection outcome. The rationale for the weighting distribution across tools is often not stated. Randall et al (2006) describe an alternative process where the scores are used to guide the selection decisions made by a panel. Performance in different competency domains across a range of tools are considered when deciding on a final rank list of applicants. Some processes include a ‘veto’ process, where assessors can raise concerns about an applicant and decision be made to ‘select out’ these applicants irrespective of their score (Gale et al., 2010). Defining how information from a number of selection tools should be combined to make selection decisions has been identified as one of the gaps in the current selection literature (Patterson, Knight, et al., 2016; Prideaux et al., 2011; Shulruf et al., 2018).

Summary and Gaps

This review of the extant literature on vocational medical selection has been unable to satisfactorily answer the research question:

What is an effective process for selecting applicants into vocational medical training programs?

The literature on vocational medical selection tells a story about different groups across the world, seeking an *effective process,* struggling with the same set of issues. There are increasing numbers of applicants for a fixed number of vocational medical training positions. Selection coordinators are faced with the challenge of selecting trainees from a pool that contains far more suitably qualified applicants than there are available positions. They must also have a method to 'select out' applicants who are not suited to the training program. There are a range of tools available to assess the suitability of applicants but information about these tools is mainly limited to their reliability and political validity. The utility of each tool is dependent on a number of variables that includes the content, scoring system, training of assessors and number of assessments. When looking for an *effective process*, the predictive validity of tools is held in high regard but gathering evidence on this is difficult. Each tool provides information about the applicants which must then be used to make a binary decision about selection. The whole selection process must be undertaken in a feasible manner with regard to the time and resource impact on both the selecting institutions and the applicants. Research on stakeholder acceptability is limited to the opinions of applicants and assessors in a selection process. Arguably, what is most needed is a framework that can be used to navigate these challenges. Whilst there is a paucity of published literature discussing specific frameworks used for vocational medical selection (Roberts et al., 2017), there are examples emerging.

Fiona Patterson has published a framework that can be applied to designing both medical student and vocational selection processes, initially in 2013 then updated in 2016 (Figure 2) (Patterson, Ferguson, et al., 2013; Patterson, Knight, et al., 2016). The model lays out the different components of a selection process showing the selection tools as one component integrated into a wider system. The process starts with stakeholder consultation and careful consideration of the desired attributes for applicants. Selection tools are then chosen and matched to assess these attributes. Evaluation and feedback mechanisms are built into the model. The principles of this model have been referenced in case reports as a valuable template on which to base a selection process (Gale et al., 2010; Goodwin et al., 2014). Roberts et al (2017) refers to Patterson's model when describing the emerging literature around selection processes based on the principles of competency based medical education. Both Patterson and Roberts call for further research into the theoretical frameworks that underpin selection processes.

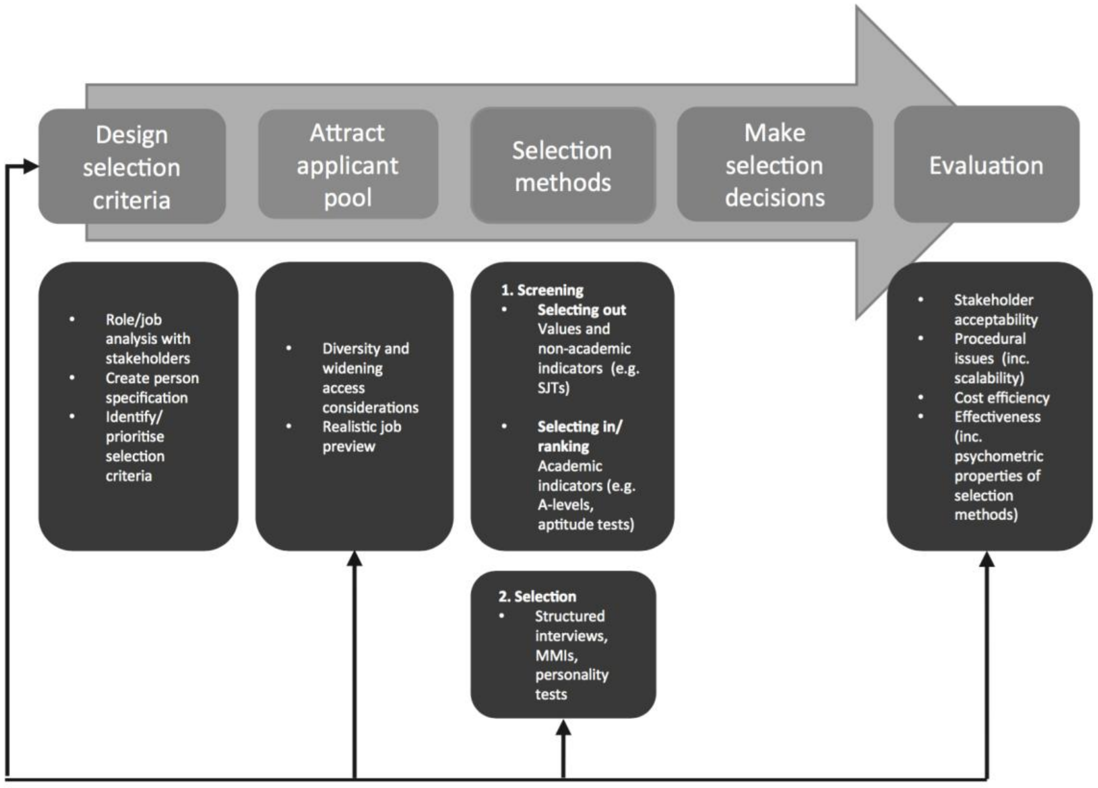

Design selection criteria
Attract applicant pool
Selection methods
Make selection decisions
Evaluation
• Role/job analysis with stakeholders
• Create person specification
• Identify/ prioritise selection criteria
• Diversity and widening access considerations
• Realistic job preview
1. Screening
• Selecting out Values and non-academic indicators (e.g. SJTs)
• Selecting in/ ranking Academic indicators (e.g. A-levels, aptitude tests)
2. Selection
• Structured interviews, MMIs, personality tests
• Stakeholder acceptability
• Procedural issues (inc. scalability)
• Cost efficiency
• Effectiveness (inc. psychometric properties of selection methods)

Figure 2: Patterson et al (2016) model for design of selection processes
(Reproduced with permission – see Appendix A)

An assessment framework has also been proposed as a model on which to base selection (Prideaux et al., 2011). Educational assessment and selection both require a process to make a series of judgements about a person, with endpoints that represent high stakes to the individual. A variety of measurement tools are available to provide information to coordinators of such a process to make a decision about progression (in education), or appointment (in selection). The shortcomings of focussing on tools and their psychometric properties has been well described in educational assessment (Hodges, 2013; Schuwirth & van der Vleuten, 2006) and the same issues are potentially seen in selection. The employment of assessment principles in selection design is still evolving.

While there is a growing literature in vocational medical selection, how this can be applied to the Australian context is unclear. The structure and governance of training programs varies across the world, and practices in other jurisdictions may not be generalisable. Since the findings of the Brennan Report were published in 1998, publications concerning the Australian context have been limited. An exception to this is the evolution of selection into general practice training which has been the subject of a number of articles regarding evaluation of selection tools, social and predictive validity (Burgess et al., 2014; Burgess, Roberts, Sureshkumar, & Mossman, 2018; Patterson, Rowett, et al., 2016; Roberts & Togno, 2011). Literature regarding other specialities indicates these medical specialty colleges are early in their journey of developing selection processes. Emergency medicine training has only recently begun moves to formalise selection (Chu, Kaider, & Johnson, 2017; Thomas, 2017). The Royal Australasian College of Physicians is still in the process of developing a formal policy for selection into training (Royal Australasian College of Physicians, 2018). The surgical specialties report to have based their selection processes on the Brennan principles however there is limited publicly available information to ascertain if tools and processes used are effective (Grantcharov & Reznick, 2009; Oldfield, 2013).

The following chapters present a mixed methods case study as an example of vocational selection which sought to understand the challenges inherent in designing and implementing a selections process, within an Australian context. The case study and the presented literature provide a basis for discussion of an alternative framework to approach vocational medical selection.

3. METHODS

The published literature on vocational medical selection presents a mixed picture, with a capacious and varied box of tools but uncertainty of how they should be used together. Coordinators of selection can access a vast range of tools, and information is available on how to strengthen their psychometric properties. The uncertainty surrounds how to use data obtained from multiple tools to reach selection decisions. A mixed methods case study approach was chosen to further explore selection for specialist training.

Case study is a research method that allows quantitative and qualitative research methods to be used together in a complementary manner, to develop new knowledge about the topic of interest in context (Merriam, 1998). Selection itself, requires elements of both quantitative and qualitative research approaches in its processes and conclusions.

Elements of selection draw on the positivist paradigm used in quantitative research. That is the view that, a fixed reality exists, it is directly observable and can be measured (Tavakol & Sandars, 2014). A positivist view in selection would be that; *among all the applicants are a defined number that are truly the best and should be selected, the challenge is finding right method to find these few people amongst the many.* This view holds that knowledge is to be discovered and cannot be socially constructed. A selection process that emphasises standardised scoring of applicants' behaviour and performance and relying on these scores to make selection decisions is consistent with this positivist paradigm.

Other aspects of selection are rooted in the philosophical worldview of constructivism, more typical of qualitative researchers. That is, the reality we experience is socially constructed, that interactions between people define the truth of the way things are. Multiple realities exist, and these are defined by people's interactions with each other and their environment. The truth is open to interpretation and unable to be measured. So in selection a constructivist view may be that; *among all the applicants, there are many who would be appropriate to be selected as trainees, for different reasons, that will be determined by the applicants' interactions with assessors, future work colleagues, patients and institutions.* Selection processes that use and value a diversity of assessors to make subjective assessments about applicants are using a constructivist paradigm.

It is fitting then that at a mixed methods approach be used to explore the research question through both positivist and constructivist lenses. Qualitative inquiry is suited to gaining an understanding of a phenomenon from the perspective of participants – in this study this was done through a group interview with selection coordinators. The positivist view of selection garners meaning from the use of statistical procedures. The quantitative approach taken in this case study was to evaluate the selection process through the statistical analysis of individual tools and associated decision-making processes. Case study methodology allows exploration of a complex phenomenon (in this case selection) in a real life context. The interactions between people, conflicting priorities, contextual challenges and resource limitations can be studied in situ as they naturally occur, in contrast to an experimental design where realities seen as imperfect can be controlled (Yin, 2009).

The Case Context

This case study explores the processes used for selecting trainees into a medical specialty training program in a single state of Australia. The state and the training program involved have been intentionally kept anonymous throughout this thesis. This decision was taken to protect the identity of the individuals and hospitals involved. Although specialty programs span the entire country, the individuals involved in selection are part of a relatively small community. Divulging the specialty and state would likely in turn identify the individuals who coordinated the selection process. This case study involves a critique of the selection process. The intention is to use this critique to explore and understand the issues faced by coordinators of selection process. The intention is *not* to criticise individuals or question their professionalism or integrity. There were concerns that identifying the case study could damage the reputation of past and future selection processes. It is also important that the content of selection tasks and marking schema are not identifiable to potential future applicants. Therefore, throughout this thesis, the case study will be referred to as the State-Wide Specialty Network (SWSN).

Like most specialty programs in Australia, the training program in the case study is governed by a national medical specialty college. While the training curriculum, examinations and awarding of fellowships is coordinated centrally by the college, the process for selecting trainees into the college occurs at a local level. Individual training hospitals, or groups of hospitals (networks), across Australia and New Zealand, devise their own processes for selecting trainees to work in their hospitals, with little specific guidance from the college. Once a doctor is successful in obtaining one of these positions within a recognised hospital, they may apply to the college for entry into the training program. Each hospital/network must make decisions about how to run their selection process and grapple with the challenges that selection presents. This case study was bounded by the selection processes used for two cohorts of applicants who applied for training positions in the SWSN. Specifically, the selection process held in 2014 for entry into training starting in 2015 (this data is referred to as the 2015 cohort) and the selection process held in 2015 for entry into training in 2016 (the 2016 cohort).

An evaluation of the SWSN process was performed using a mixed methods approach. First, qualitative inquiry was undertaken through a group interview with selection coordinators. This case study afforded the opportunity to explore the motivations behind the development of a selection process, the rationale for the choice of tools and modifications made to how they were used in each cohort, and also to investigate the logistical issues and challenges that arise when administering high stakes selection. A quantitative analysis was performed using descriptive statistics to examine each selection tool used in each cohort. Since there was variation in the format, content and scoring rubrics for the same tool type between the two cohorts data from the two different cohorts was used to compare the utility of different selection tools. Quantitative analysis included modelling alternative combinations of tools to look at the influence on selection outcome. Finally, the social validity of the SWSN process was considered through the evaluation of exit surveys completed by the applicants for the 2016 process.

The Researcher

The researcher and author of this thesis is a clinician and educator working within one of three hospitals involved in the SWSN. I was not directly involved in the selection process in this case study and do not currently have a role in vocational selection. I coordinate selection processes for pre-vocational doctors in our hospital. I am interested in constructing a selection process that leads to us employing high quality doctors, is efficient in terms of time and resources, and is fair to all applicants. The SWSN is an accessible case study that allows exploration of selection in detail. I conducted a group interview with two selection coordinators who work within the same hospital and was one of two researchers who coded the interview transcript.

Sources of Data

Three main sources of data were available to undertake this case study; documents, a group interview and raw selection data for the two cohorts. A series of documents used in both cohorts were used to gather information about the tools and processes used (Table 1). These include information given to applicants, descriptions of the selection tools used and the associated marking criteria. The documents provided useful information for triangulation with both qualitative and quantitative results. A semi structured group interview was held with selection coordinators to explore their perspective on the process. The transcription of this interview was used for thematic analysis. Deidentified raw scores for applicants in 2015 and 2016 cohorts were used to undertake the quantitative analysis of tool performance and examine the impact of alternative algorithms for combining scores

Table 1

SWSN documents

Document Title	Appendix
Example referee report form	B
2015 Marking Guide	C
2015 Advertisement training positions	D
2016 Interview questions marking guide	E
2016 Shortlisting marking guide	F
2016 Application questions in lieu of cover letter	G
2016 Applicant exit survey	H

Group Interview

The purpose of the group interview was to view the selection process through the lens of selection centre coordinators. Of specific interest were the rationale for the design of the selection process and logistical issues.

A face to face semi-structured group interview was held on February 10th, 2015 between the researcher and two selection coordinators from the SWSN. The selection process model described by Patterson et al (2013) was used to structure the questions to gather information about each stage of the selection process. Prior to the interview documents from the 2015 and 2016 process had been reviewed by the researcher to form an understanding of the selection process. The interview provided an opportunity to clarify information as well as seek to understand the motivations of selection coordinators, the challenges and successes they experienced, and the insights they had regarding selection. The interview allowed the researcher to probe background reasoning as to why decisions were made in the design of the selection process.

The interviewees were both involved in the design and implementation of the annual selection process over several years including the two cohorts that are part of this case study. Interviewee 1 was a manager with experience in human resources and junior doctor recruitment. Interviewee 2 was a clinician with experience as a director of training and knowledge of the medical training culture throughout the network. They were interviewed together as each had specific areas of knowledge that would allow a complete picture of the selection process to be garnered.

The interview transcript was analysed by two researchers using NVivo software (NVivo, QRS International) using a basic qualitative approach (Merriam, 1998). The initial analysis used open coding identifying categories and emerging themes. Researcher one (the author) is a clinician working closely with trainees. Although not directly involved in recruitment for training positions he had been through selection and training himself and had been a referee for applicants during the 2016 process. The interviewees were also work colleagues of researcher one. There was the potential that his own experiences would influence his analysis of the interview. The second researcher was chosen to analyse the interview to balance the potential of this 'insider knowledge' influencing the analysis. Researcher 2 was not involved in specialty selection or training and was able to provide a fresh outsider viewpoint when considering vocational selection, enhancing the reflexive process.

Post hoc Statistical Evaluation of Tool Performance

A thorough appraisal of the SWSN process had not been performed prior to this study. Therefore, an independent evaluation was undertaken to explore the utility of tools and decision-making processes. Information on the predictive validity of a selection process is useful to evaluate effectiveness but this was not possible with the information available. At the time of this study, trainees had not yet reached the stage of their training where they undertake exams. Further information from work-based assessments is not collated in a systematic way. Therefore, there was no quality information available on the subsequent performance of the trainees selected and none on the applicants who were not selected. Consequently, the evaluation used the data available from both cohorts to review three components of the selection process: selection tool performance, how selection decisions were made, and the social validity of the process.

Selection Tool Performance

Descriptive statistics were used to consider applicant's scores across the selection process. For each tool, the descriptive indices used were the range, mean and standard deviation of scores. The inter-rater reliability was calculated for the interview questions that used multiple assessors. As the interview questions were scored using continuous data, intra-class correlations were used to measure consistency of scoring between assessors (Cook & Beckman, 2006; Laschinger, 1992). The contribution of each tool to overall selection outcome was also calculated as a percentage of the total final score used for ranking for selection. This allowed consideration of the contribution of each tool to selection outcome, both in terms of weighting of the score and discriminatory value in differentiating between applicants. Other statistical methods were considered to evaluate the data including factor analysis to look for tools that potentially tested the same constructs, and also discriminant analysis to examine the ability of each tool to differentiate applicants. After consultation with a statistician it was advised that the cohort size was too small for meaningful use of these tests.

Making Selection Decisions

It still remains unclear from the published literature how different selection tools should be combined to reach selection decisions. In the SWSN process, scores from tools were summed and a ranked list of applicants developed. Final selection decisions were made following the rank order of this list until all training positions were filled. To evaluate the impact of weighting and combining scores on the selection process outcome, alternative combinations of selection tools were modelled.

Cohen's Kappa statistic was used to measure agreement between the selection process used by SWSN and a number of possible alternative combinations of tools (Table 2). Cohen's Kappa is used to compare a gold standard measurement method to an alternative (Watson & Petrie, 2010). Since, there is no gold standard selection method, alternative combinations of selection methods were compared to the current practice. The binary outcomes of 'Selected' or 'Not Selected' were used. There was a fixed number of positions available for training in each year (15 in 2015, 14 in 2016). To be selected, an applicant must have ranked in the top 15 or 14 places of their cohort.

Perfect agreement is achieved when Cohen's kappa equals 1. The following levels of agreement are generally accepted when judging levels of agreement (Watson & Petrie, 2010):

Poor if κ <0.00

Slight if 0.00 $< \kappa <$ 0.2

Fair if 0.21 $< \kappa <$ 0.4

Moderate if 0.41 $< \kappa <$ 0.6

Substantial if 0.61 $< \kappa <$ 0.8

Almost perfect if $\kappa >$ 0.8

For each of the two cohorts, alternative selection models were compared to the current process. The rationale for each alternative is given in Table 2.

Table 2

Alternative selection models

Alternative Combination of Tools		**Justification**
1	Traditional Method: (CV, +/- cover letter, interview referee report)	Historically this approach was used for many years. Does the current approach achieve a significantly different result?
2	Shortlisting process only: Pre-selection centre items	Selection using only information from instruments used prior to the selection centre. 2015: CV, Cover letter, Personal Statement 2016: CV, Four Written Questions, Referee reports To investigate if there is justification for running a selection centre at all. If a similar result comes from using tools supplied by the applicants, then there may be an argument to not use a selection centre saving considerable time and resources.
3	Selection Centre Items only	Is there justification in only using information gained from the selection centre to make selection decisions and using the pre-selection centre tools only for shortlisting?
4	SWSN Process without referee scores	Use all tools in the SWSN process with referee reports scores removed from the overall score. Referee reports have been considered an unreliable instrument to use in selection. This will assess what influence they had in the overall results.
5	SWSN Process without interview scores	Use all tools in the SWSN process with interview scores removed from the overall score. Interviews are the most time and resource intensive instrument used. Do they significantly alter the result?

Social Validity

Stakeholder acceptability is an important consideration when evaluating the effectiveness of a selection process. The political validity describes the extent to which stakeholders view selection as appropriate (Patterson, Ferguson, et al., 2013). Social validity extends this concept further to include perceptions of fairness, transparency and the opportunity to display skills and knowledge (Schuler HI, 1993). Applicants in the 2016 cohort completed a paper-based exit survey (Appendix H) (before they knew whether they had been successful). Applicants were asked to state their level of agreement with a number of statements including:

- *This task seemed to be an appropriate way to make decisions about selection into (specialty) training.*
- *I had the opportunity to demonstrate my skills and abilities.*
- *This task was a fair assessment.*

A four-point Likert scale was used to capture responses. The neutral option was omitted to force agreement or disagreement. Applicants were also invited to comment in free text responses.

Survey responses were studied to explore the social validity of the case study.

Summary

This case study uses interview, SWSN documents and data from two cohorts of applicants to explore selection into vocational medical training. The group interview provided the opportunity to view selection through the lens of selection coordinators and understand the motivations and challenges behind the design and implementation of a selection process. The findings from the interview are presented in Chapter 4. And independent review of quantitative data was undertaken to explore selection tool performance, decision-making processes and applicant acceptability of the selection process. This evaluation is presented in Chapter 5. This study received ethics approval from the Social and Behavioural Research Ethics Committee Project number 7075*.*

4. THE STORY OF THE STATE-WIDE SPECIALTY NETWORK (SWSN)

This chapter outlines the selection process used by the State-wide Specialty Network (SWSN). An overview of the processes used in two different cohorts of applicants is described. The analysis of the group interview with selection coordinators is then presented. The purpose of this chapter is to explore the reasoning behind the construction of this selection process and also the rationale for the choice of selection tools and decision-making processes. Understanding the motivation of selection coordinators and the challenges they faced can help inform how effective selection processes could be designed.

Background

The SWSN comprises representatives of the hospitals across the state that employ specialty trainees. In the past, doctors interested in pursuing training in the specialty could apply to a number of different hospitals and if successful in obtaining a position, would then apply to the college for entry into training. Each hospital had a separate selection process and many applicants applied to more than one hospital. Successful applicants would then work across a number of different hospitals during their training. The SWSN was formed to streamline this process for both applicants and hospitals and in doing so there was a complete redesign of selection practices. This case study considers two cohorts of applicants for specialty training using the SWSN selection process (2015 and 2016). In this thesis, the term *past processes* refers to selection practices prior to the formation of the SWSN, and the *current process* is the selection centre approach used in both 2015 and 2016.

In each cohort a shortlisting process was used to identify applicants who were invited to participate in a number of further selection activities at a selection centre. In 2015, there were 29/69 applicants shortlisted to attend the selection centre for 15 available training positions. In 2016, there were 32/79 applicants similarly shortlisted for 14 available training positions. The selection centre involved undergoing an interview and completing a range of other selection tools. Each tool was scored and the sum of these scores formed the final selection score which was used to rank applicants. Training positions were offered to the top 15 and 14 applicants in 2015 and 2016 respectively. Figure 1 provides an overview of the selection tools used and their weighting in both cohorts.

The shortlisting and selection centre processes differed slightly between years. Descriptions of each tool, content and marking criteria, are presented in Chapter 5. This chapter explores how selection coordinators constructed the selection process through analysis of the group interview.

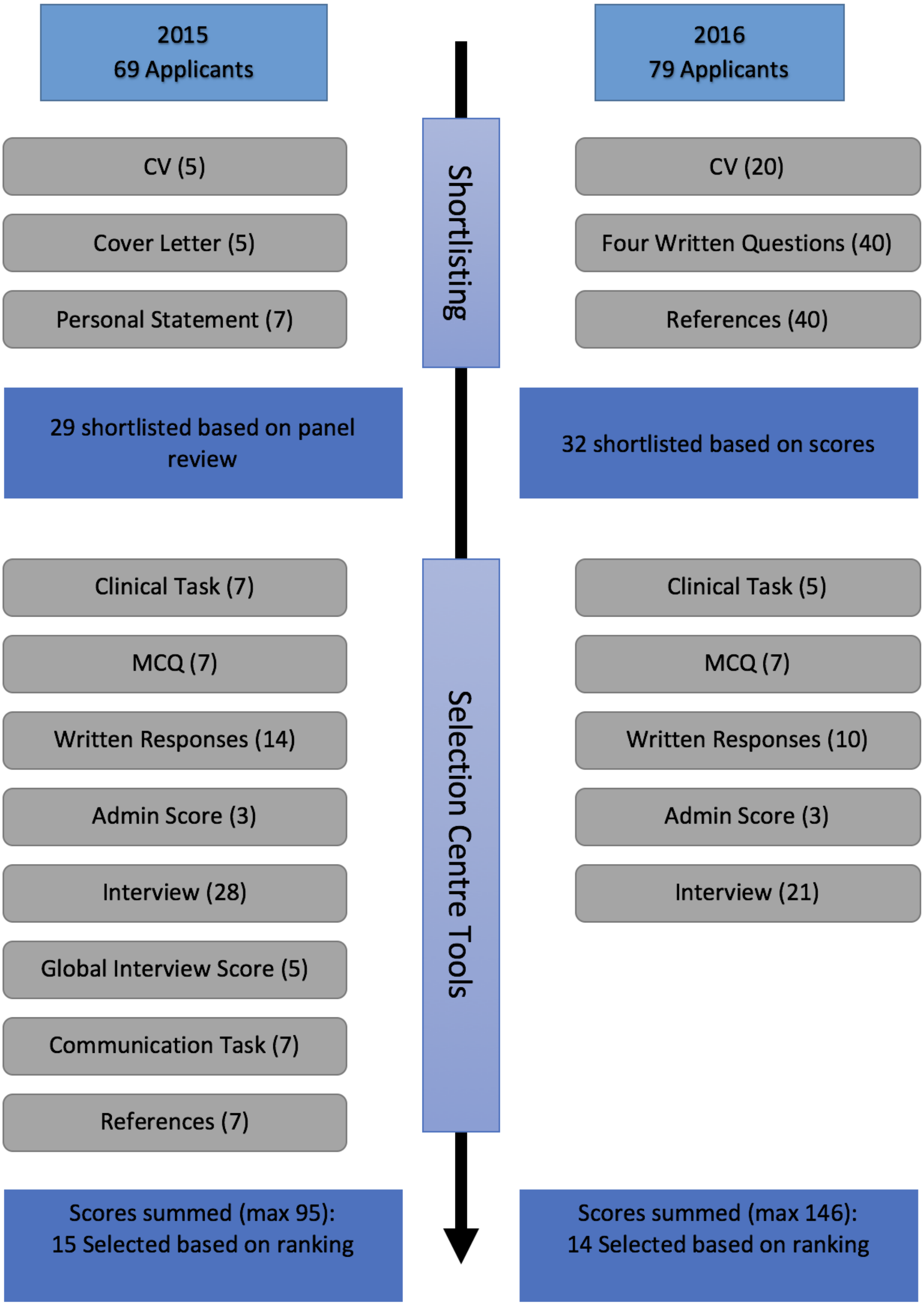


Figure 1: Summary of selection processes
The maximum score available for each item is denoted in parentheses

Group Interview Themes

Several themes emerged from the analysis of the interview transcript that provide a useful framework to discuss the SWSN process. Researcher 1 read and coded the interview into nine themes. Researcher 2 used open coding to classify information into eight broad categories with several subcategories. The researchers met to discuss themes and discovered multiple areas of commonality. On review of the researchers' coding, seven common themes emerged. Table 1 summarises the initial themes from each researcher and how they grouped to form common themes.

Table 1

Summary of themes from the group interview

Final Themes	Researcher 1 – initial themes	Researcher 2 – initial themes
Dissatisfaction with the selection process	Dissatisfaction with process	Concerns
Purpose	Aims of selection process	Purpose
Fairness	Fairness	Concerns: Fairness
Unease about inexact science of selection	Inexact science of selection	Concerns
Informality and Pragmatism	Informality	Informality
	Opinion	
	Pragmatism	
Improving the program	Quality improvement	Evaluation
Tools and Decision-making	How decisions were made	Process
		Development of selection process
		Tools Used

Overall, the selection coordinators had concerns about both the past and current selection processes. They described past processes as lacking rigour but remained worried that the current process too was imprecise. They were clear about the purpose of the selection process and the reasons for any modifications made between cohorts, which were centred on fairness for applicants. They placed emphasis on the selection tools chosen and discussed their efforts to modify the content and format to improve utility. Throughout the interview both researchers observed that the selection coordinators were motivated to improve the selection program and looked to the literature for guidance on how to do this. When clear guidance could not be found, they used pragmatism and instinct to make decisions about the design process. Anxious about ensuring the process was fair, selection coordinators valued the scoring system, which they viewed as more objective, over qualitative ‘subjective’ assessments made during selection.

Theme 1: Dissatisfaction with the selection process

The interviewees expressed dissatisfaction with both the selection processes of the past and also elements of the current process that they were working to improve but were yet to find a satisfactory solution. The past process was perceived as having flaws and this was the motivation for changing to the current process. Previously, there had been a number of separate processes occurring with disparate selection criteria and methods.

> *…if we go back a few years there were …(many)… selection processes going on (in the State), one at (Hospital X), and one at (Hospital Y) ... And they were actually operating under different criteria and different selection methods, even though each of them had very basic methodology for selection. (Interviewee 2, p1)*

There was anxiety about getting the process right and acknowledgement that designing and implementing a selection process was difficult. The shortcomings of various available tools were acknowledged. Coordinators were concerned about the accuracy of CVs and referee reports but still used these tools.

> *I read one article where, I don't think it was in the health setting, but it was a professional industry, I forget which one, they had something like a 12% fraudulent CV rate for people going for professional positions. So they went to all the trouble of actually, you say you went to this school, this uni, did this course, that sort of thing, and they contacted all of the agencies involved and some of it was simply not true. (Interviewee 2, p12)*

They were uneasy about applicant interviews, outlining their fear that good interviewees do not necessarily make good trainees even though (they noted), they had no evidence to qualify this opinion.

> *…if you can fire off the first thing that comes to your mind and maybe sound convincing and fluent at a panel interview … then is that really what we want? Or do we want people who can pause, think and give a concise, considered answer to questions? Now, trying to ask people to do that in a panel interview is virtually impossible, so we said, "Right, well we should look at other tools to include in the selection process." And so then we had to decide what those tools would be.* (Interviewee 2, p2)

The coordinators were aware of the complexities of selection. At times they made attempts to address these challenges by modifying the way tools were used (e.g. removing tools from the process that they deemed not useful). In other instances (e.g. with referee reports) they acknowledged the flaws of the tool and used it anyway. It was evident they were anxious about the conflict between their instincts, their expectations of what should be in a selection process, and the available evidence.

Theme 2: Purpose

The selection coordinators had a clear purpose and made decisions in the process design that reflected this. A number of purposes were referred to including; identifying the attributes of suitable applicants, using the most appropriate tools to assess these attributes in applicants, identifying applicants who would be successful trainees in the first few years of training (rather than those who would make competent specialists). Desired attributes for trainees included a commitment to completing the tasks required for successful training, a solid work ethic and sound communication and reasoning skills. They referred to a medical specialty college document, that defined a list of desirable qualities for trainees.

> *So I think what we wanted was people with the capacity to be a successful trainee but also the attitude to match, that they would knuckle down and take their learning and training seriously…good reasoning, problem solving, situation or judgmental skill. (Interviewee 2, p2)*

Interviewee 2 differentiated between the attributes needed in specialty training and those skills needed to be a competent specialist. He spoke against trying to select for competent specialists, as the pathways to specialisation are varied and there is much to learn in the training journey.

> *I think we are selecting for success in basic prep. Yeah, because trying to mould people to a certain mould [laughs] you know, when they're finally through is (a) not possible, and (b) probably not desirable. So there's many paths to tread, especially in advanced training. (Interviewee 2, p5)*

The purpose of redesigning the selection process was also about improving the quality of trainees selected. There were concerns about previous trainees who were unable or unwilling to attempt and complete examinations in a timely manner (trainees are expected to sit their exams in the third year of the program, but this was being postponed by trainees who felt they were not yet ready, delaying their progression through training which impacts on the hospital's ability to fill advanced training positions). However it was unclear how the new process would improve the quality of trainees and avoid such trainee issues.

> *…in the previous years when we had insufficient quality applicants, basically scrounging, so the people entered training but did not take it seriously, were not progressing, not preparing for college examinations would be the best example, or sitting the examinations and failing badly. But the biggest problem was that people were just deferring often multiple years, and just basically occupying a job, a position. (Interviewee 2, p2)*

While the desired attributes were readily identified, the links between these and the selection tools used were not overtly clear. There was no formal documentation of attributes, competencies or domains, nor any reference to these during the final decision-making process. It was not clear to the researcher whether coordinators thought their purposes had been achieved by the selection process.

Theme 3: Fairness

Throughout the interview, the selection coordinators emphasised the importance of having a fair process and discussed their anxieties about ensuring procedures and decision-making were transparent and just. In the last few years there had been an expected increase in applicant numbers. In contrast to past experience, when there had been a shortage of applicants, entering training had now become highly competitive. When designing the process, the interviewees wanted trainees to be selected in an impartial, non-discriminatory way. They talked of how they made efforts to research best practice in selection and attended workshops to gather information that would help them.

> *So with that change* [increase in applicant numbers] *we felt we had a responsibility to make the selection process meritorious and fair. (Interviewee 2, p1)*

Fairness was also a central tenet when considering what information was communicated to applicants prior to and during the selection process.

> *I1: … so we've included a paragraph that reads along the lines of, "You've shortlisted (sic). You'll be required to undertake several tasks which may include answering clinical centre questions, MCQs, communication station."*
>
> *I2: I think that's very fair*
>
> *I:1 So this was an inclusion to try and make it more fair, and for them to understand the importance of the content in their CV and covering letter.*
> *(p6-7)*

A number of changes were made to the selection process between 2015 and 2016 and often the rationale used was to enhance fairness. Fairness was the justification for having an *increased* number of assessors so that an average score could be used (in interviews). Fairness was also the justification for *reducing* the number of assessors to ensure consistency across the cohort.

> *So like everything else we didn't want marker bias or in turn marker bias, so for '16 I think … all the applicants were scored for a particular task by one scorer. So if that person was biased, at least he was biased for everybody. (Interviewee 2, p16)*

Consistency across markers was viewed as enhancing fairness for tools like written questions while for other tools incorporating diverse opinions of the same applicants was seen as countering bias and was therefore fairer. A panel was used for interviews and an average of the scores from each panel member was used for balance.

> *... so after the first candidate we'd talk about how we thought that person rated to ensure that we are of the same understanding. But also, the reason we develop an average for each question is so that if I'm a dove and (Dr X's) a hawk… (Interviewee 1, p25)*

This desire to be fair and consistent when marking was also the justification for exclusion of a tool when agreed procedures to ensure fairness could not be achieved during the 2016 selection process.

> *The communication task, I think that we actually took that off the final score because we had some resource issues whereby we had to, at the last minute we had sick leave, and so the people marking the communication station were actually different, and it was obvious when we reviewed their marking…that they weren't marking… So we actually took that off… (interviewee 1, p16)*

The *perception* of fairness was also central to the decisions made about both selecting trainees and choosing selection methods. This was seen as important for justifying the selection process to others and demonstrating the integrity of the process when giving feedback to unsuccessful applicants.

> *And just thinking about that, if you've got 14 positions and you've done something that ranks people, and you take the top 30 for 14 positions it sounds pretty fair. (Interviewee 2, p12)*

It was clear their intention to have an assessment process that was just, impartial and transparent.

Theme 4: Unease about the inexact science of selection

Selection coordinators wanted to create a fair and robust process to select the best trainees but were uncertain about how to achieve this. They expressed their concerns about past processes and voiced consternation about not choosing the best trainees. They were worried that the people who performed well at interview may not necessarily be the best applicants because performance at interview may not reflect ability as a trainee. When it came to finding a solution to this dilemma they were not clear about how to proceed.

> *Yeah, I think I'd have to look at the research science around it a little bit more. I've looked a bit; it certainly confirmed my impression that it's pretty inexact. (Interviewee 2, p21)*

The coordinators looked to the published literature, other colleges and other industries for guidance on how to design the process. They were surprised that it was difficult to find a clear answer about how they should conduct selection.

> *So we did do some looking at literature and trying to learn a little bit about what made selection valid; that was quite interesting. It probably reinforced some gut feelings; a good example there would be looking at somebody's CV and conducting a panel interview, only it is in any other industry completely insufficient and there's been some studies done on the outcome of that sort of selection, and the outcome is poor. (Interviewee 2, p1)*

Even though there was lack of clarity, they still had to progress with selection, albeit with disquiet about the tools they were using. An example is the discomfort with using referee reports but still including them despite reservations:

> *…how people tick boxes on rating scales is completely inexact, but should we pay any attention to referee reports, no attention, have them, not have them? So at this stage we've decided to still include them. (Interviewee 2, p19)*

Despite referee reports being 'inexact' they were given a high weighting in the 2016 process (27.4%). This decision indicates a degree of inertia at play: referee reports had always been used, and without obvious superior alternatives, they continued to be included.

Similarly, they had reservations about interviews, but these were a significant contributor to the selection process. Coordinators were worried that interviews, while a good assessment of communication skills, might not be a good indicator of type of trainee they sought. They wanted to examine skills other than communication and expressed the view that tools other than the interview would be needed;

> *Panel interviews are quite good at seeing if somebody can communicate, so communication; and some knowledge about professional skills other than just clinical knowledge, although we did add some tools that probed clinical knowledge to a superficial degree. (Interviewee 2, p2)*

> *Also when we're trying to select doctors into a (specialty) training program I've personally always harboured this horrible feeling that the best performers at a panel interview may be not the sort of people we want training as (specialists). (Interviewee 2, p2)*

Despite these concerns the interview contributed 29.5% of marks to the final selection score in 2015 and 14.4% in 2016. Again, there is a sense of unease about the use of a tool but without alternatives persistence in using it for selection as is clear in the above quotes. The lack of clear evidence in how to conduct selection led coordinators to take a somewhat practical approach to the design and conduct of the selection process.

Theme 5: Informality and Pragmatism

Acknowledging selection processes were imprecise, but were still needed, a pragmatic approach was often taken. When deciding how to use tools they made decisions in a variety of ways. On one hand, they followed the practice of authoritative bodies and considered what specialty colleges were doing and based some aspects of the selection process on this. In other circumstances, where there was no established best practice, they reverted to using their own judgement and rationale.

> *Yeah, a sort of educated guess, I would say, as to how we should weight it. (Interviewee 2, p24)*

Although they expressed a preference for using scores and linked this to fairness and objectivity, there were comments by both interviewees that referred to the role of using ‘gut feelings’ and individual ‘impressions’ when making decisions:

> *But that year (2015) we didn’t have a scoring guide to mark the applications. So we say, each took a pile of 15 applications each, reviewed them, made comments and shortlisted in our minds. (Interviewee 1, p8)*

This pragmatic approach to assessing applications also extended to their own internal review processes. There were a number of examples where changes were made to the process based on a qualitative review, with no formal evaluation taking place. The communication station used in 2015 was dropped from the 2016 process because the scoring system used was deemed to be too inconsistent. The cover letter in 2015 was replaced by short answer written questions in 2016 because the impression was the cover letters did not provide valuable information. In these two examples it was felt that the changes resulted in improvement. There was no quantitative review of tool performance to inform their thinking but rather a reflective, qualitative appraisal of how the tool functioned that led to the changes.

In designing the selection process, the selection coordinators drew on their own experience from many years selecting trainees as much as evidence from the literature. So while the selection coordinators were keen to be guided by evidence, they were often compelled to design a selection system, that to them had face validity, based on their own experiences.

Theme 6: Improving the Program

The selection process was evolving, with tools and procedures continually being modified with the aim of improving the program. The selection coordinators made reference to efforts that had been made to review the selection process, however no structured evaluation had been done to evidence quality improvement strategies.

> *An important part of this process for the last few years, including the two years of interest, is that when the process has finished we do pretty formally review the process, so we'll do a kind of team session to see whether we thought it went well, what went well, what didn't, what should we change and then decide on the changes. (Interviewee 2, p10)*

They acknowledged the difficulties measuring the success of the selection process, that is, the predictive validity. They reported that work based assessments used for trainees are poor discriminators of performance.

> *We've discussed trying to do some sort of correlation with… the people we've selected; can we measure their performance in some way, to see if we can confirm that we've selected high performers. But when you delve into the methodology that's virtually impossible to do. (Interviewee2, p10)*

Efforts were made to review and modify the selection process between the two cohorts. This included the removal of the *Global Score* for the interview, changes made to the clinical task and modifications made to marking guides. The reasons for the changes seemed to be based around the themes already identified: greater fairness for applicants, to ensure a pragmatic and efficient process and to improve the perceived accuracy of the tools and process.

The financial cost of the process had not been considered formally. When prompted to discuss cost by the researcher and both interviewees responded that it was difficult to quantify or even to plan how to quantify the cost of the process.

> *It's really difficult to estimate the time. (Interviewee 1, p28)*

> *There's not a neat start and finish either. Well the start's fairly neat, but the finish is definitely not neat. (Interviewee 2, p29)*

The resource implications were seen as a lesser issue compared to their motivation to have a fair and robust process.

> *The personnel involved here are all paid to do their role, whether we do this* (process) *or not. So I'd say the cost is zero; there's no additional cost… Whether we do it or not there's not a monetary cost to the health system. (interviewee 2, p28-29)*

While there were efforts to improve the overall selection process, there were no specific training or standardisation procedures for assessors. Each year, all the assessors in the selection process were involved in the design and content of selection tools. A panel of four specialists and the selections manager met in person and shared information regarding tools and questions via email in the lead up to selection. However, tools were not piloted and there were no calibration exercises performed.

Theme 7: Tools and Decision-Making

A major theme of the interview concerned the quality of selection tools and how these tools were used to make selection decisions. The need to have a discriminating spread of scores was the motivation for increasing the number of tools.

> *Yeah, so the more items you have the more the score just stretches right out [laughs], and that's a statistical thing, mathematical thing. (Interviewee 2, p15)*

This illustrates the pragmatic approach, making design decisions that primarily facilitate the task of differentiating between applicants.

There was particular concern about the ranking process and ensuring the cut-off scores for selection were meaningful.

> *…because somebody has to be the last applicant that gets a position, and somebody has to be the first applicant who doesn't get the position. And so I was quite anxious that that cut-off had some validity to it and some separation that you could point to this person, say they scored significantly higher than this person who was the first one to miss out. So that was quite important. Trying to pick the really top, the first person to be offered a position is easy, and the bottom person who you don't really want to employ ever, even if you had a vacancy is easy. But that middle ranking cut-off point, very, very tricky. And so it was anxiety around that that drove the whole process. (Interviewee 2, p14-15)*

The coordinators had a preference for the use of numerical scores over qualitative judgements for both the selection tools and determining final selection outcome. When asked about the shortlisting process in 2015, there was discomfort about the manner in which this was done. Cover letters and CVs were reported to be reviewed and shortlisted based on an overall subjective 'impressionist' assessment of the applicant.

> *Look, to be honest, it was based on our impression both individually, and then as a group, an impression of the quality of the application, the info and the applicant. So it was bit sort of impressionist. (Interviewee 2, p9)*

Uncomfortable with this subjective assessment, they applied a scoring system to these items the following year in 2016. Scoring was seen to provide objectivity and psychometric rigour to the process, in contrast to tools viewed as impressionist or qualitative

> *Well, I think my opinion, it's only an opinion, I haven't got much evidence for that, but my opinion is that to get an impression one is tempted to skim material. To actually score it tends to force function, reading it in much more detail, reading it and then considering it, and trying to come up with a score. (Interviewee 2, p9)*

Even when scores were used there was some unease when qualitative judgement was being used to generate these scores.

> *We dropped that* [Global Score] *because I personally, I don't know how you feel about it, I wasn't confident that it was…necessary, but similarly with concerns about this admin score it's based on my own impression and… so someone might, I guess it is an element where people could discriminate or could provide a score that we shouldn't be including in this selection process. (Interviewee 1, p27)*

Even though the selection panellists could discuss the applicants when scoring them, it was the sum of the scores that determined ultimate selection decisions. There was no option for coordinators to make final selection decisions based on review of individual tool scores and consensus opinion. The use of scores and the *absence* of a review process was thought to enhance fairness.

> *There's no secret secondary process where, "Oh my God, how did we select that person?" [laughs] … This whole effort is to stop that. (Interviewee 2, p30-31)*

This comment illustrates how the coordinators were prepared to accept the numerical value applied to tools used as the basis for selection, even if their global subjective opinion of the applicant may be that they were not a desirable trainee. Qualitative judgements were viewed negatively and described in pejorative terms.

> *Well, and that was, I'd hate to use the term "trial and error", but it was a little bit because in the year before we had used I guess an impressional (sic) score, didn't we? (Interviewee 1, p19)*

Chapter Summary

The group interview permitted an in depth look at a selection process into vocational medical training. The seven themes identified revolve around a desire for a fair and effective process, and the development of a process aimed at achieving this. Efforts to enhance the effectiveness of the selection process focused on the tools and how they were used. Modifications to the process were made based on selectors' own instincts around fairness and streamlining logistics, rather than formal review of their data. A preference for numerical scoring systems rather than using global human judgement in selection assessments was evident throughout the interview. The scoring system was viewed as being more objective and provided a means to differentiate applicants through a ranking system.

This chapter has been primarily concerned with issues around design and process in selection. There has been no formal evaluation of the SWSN selection process that takes a psychometric viewpoint or specifically investigates the utility of the selection tools and their contribution to final selection outcome. Chapter 5 reports a quantitative evaluation of the SWSN selection process initiated as part of this study. A discussion combining the design concerns and the evaluation is explored in chapter 6.

5. EVALUATION OF THE STATE-WIDE SPECIALTY NETWORK (SWSN) SELECTION PROCESS

This chapter presents an evaluation of the SWSN selection process and focusses primarily on quantitative data from both the 2015 and 2016 cohorts. In Chapter 4, the interview highlighted the selection coordinators' concerns about the quantitative rigor of the selection tools. Their appraisals of the process quality were based primarily on reflective discussions between each other. Their review was built around impressions of how well tools discriminated between applicants, as well as consideration of the logistics of administering particular tools. The opportunity is taken in this chapter to provide an independent review of the available data to explore the individual tools and the process as a whole, to understand the factors that make a selection process effective.

In this chapter the evaluation is divided into three parts:

- Selection Tool Performance. The tools used are described in terms of their content, assessment criteria, scoring and contribution to the final selection score.
- Making Selection Decisions. A summary of the decision-making process to determine selection outcome is presented and alternative models for combining selection tools are explored.
- Social Validity. Exit surveys from applicants are examined to report on the social validity of the selection process.

Selection Tool Performance

For both cohorts a similar range of selection tools was used. There are, however, some small and important differences which are highlighted throughout this chapter. The summary descriptions below are based on documents supplied by selection centre coordinators, triangulated with information about the tools and decisions made provided during the interview and in follow up email correspondence. For each tool, the format, content and scoring process is described. Table 1 summarises the performance of each of the selection tools in terms of the range, mean and standard deviation of scores for each selection tool. Percentages are presented for the mean raw scores to allow comparison across both cohorts where raw scores used are different.

Table 1

Summary selection tool data 2015 and 2016

	2015 n = 29					2016 n = 32				
Selection Tool	**Max Score**	**Range**	**Mean (SD)**	**Mean as %**	**Contribution to total Score (Weighting %)**	**Max Score**	**Range**	**Mean (SD)**	**Mean as %**	**Contribution to total Score (Weighting %)**
Curriculum Vitae	5	3.3-4.9	4.02 (0.44)	80	5.2	20	11-17	14.06 (1.48)	70	13.7
Personal Statement	7	3.4-5.9	4.73 (0.67)	68	7.3	-	-	-		-
Cover Letter	5	3.1-4.6	3.91 (0.34)	78	5.3	-	-	-		-
Application Questions	-	-	-		-	40	19-30.5	24.05 (2.29)	60	27.4
Referee Reports	7	4.5-7	5.78 (0.67)	83	7.3	40	27.5-37	32.86 (2.49)	82	27.4
Clinical Task	7	2.5-7	5.57 (1.16)	80	7.4	5	4-5	4.97 (0.18)	99	3.4
Multiple Choice Questions	7	3-7	5.22 (1.08)	75	7.4	7	4-7	5.5 (0.88)	79	4.8
Written Response Questions	14	2-13	8.66 (3.17)	62	14.7	10	4-8	6.06 (1.34)	61	6.8
Admin Score	3	2.5-3	2.88 (0.22)	96	3.2	3	1.5-3	2.734 (0.48)	91	2
Interview	28	13.4-24	20.05 (3.03)	72	29.5	21	9-19.75	14.06 (2.61)	67	14.4
Global Interview Score	5	2.4-5.3	3.987 (0.65)	80	5.3	-	-	-		-
Communication Task	7	0-7	3.79 (1.68)	54	7.4	-	-	-		-

Shortlisting Process

The shortlisting process differed between each cohort in terms of the types of tools and how they were utilised. In 2015, the decision to shortlist was based on an application that included three items; the curriculum vitae (CV), a cover letter and a personal statement. Applications were *not* scored during the shortlisting process. The decision to shortlist applicants was made by group consensus based on an overall impression of the application. Only applicants who were successfully shortlisted had their CV, cover letter and personal statement scored using a designated marking schema. In comparison, in 2016 three shortlisting items were scored for *all* 79 applicants. These were the CV, two referee reports and four written responses to questions relating to professional issues. Scores for these items were summed, and a rank list developed, with the top 32 ranked applications invited to attend the selection centre. Selection coordinators perceived the quantification of their assessments using scores in the later cohort enhanced the rigor of the shortlisting process.

Curriculum Vitae

The selection coordinators had raised concerns about potentially fraudulent claims made in CVs but used them in both cohorts nonetheless. Marking criteria required evidence of: academic excellence, participation in professional development, commitment to the specialty and awareness of the personal and social impact of disease (Appendix C). For 2015, each CV was marked by four assessors out of a maximum of five marks and the average score used to contribute to the total score for the selection process. In 2016, criteria were similar, however CVs were marked by a single assessor with a revised scoring system with a maximum score of 20 (Appendix F). The larger maximum score in 2016 resulted in the CV being weighted to contribute 13.7% to the final selection score whereas this contribution was 5.2% in 2015. The increased weighting was a direct function of the revised scoring system.

Other Shortlisting Items

A personal statement and cover letter were used in 2015 but were replaced by four written response application questions in 2016.

For the personal statement the instructions to the applicants were: *'In addition to your standard covering letter please ensure that you include a reflection on your biggest professional failure, be it clinical or non-clinical. Explore how you managed that situation and what you have learnt from it.'* (Appendix D) There are no instructions further specifying the content or length of these items. Marking criteria required applicants to show evidence of reflection on the causes of their failure and strategies they have put in place the avoid it occurring again. There was a reasonable spread of scores with a mean of 4.73 (SD = 0.67) out of a maximum seven marks.

Applicants were required to submit a cover letter but there were no instructions given to applicants regarding what to include. The marking criteria provides descriptions for grading between one and five marks: from *'Poor. Absent cover letter or very short (i.e. 1-2 sentences). Poor editing (misspelling/grammar errors). Sentences are generic'* through to *'Exceptional. Cover letter has no errors. Covers more aspects than expected in a concise manner. Very well written. Evidence included for all abilities.'* (Appendix C). Applicants needed to make assumptions about what should be in a cover letter that matched the undisclosed expectations of the marking criteria. It seems many of them did as most applicants scored around four out of five marks for this item (Figure 1).

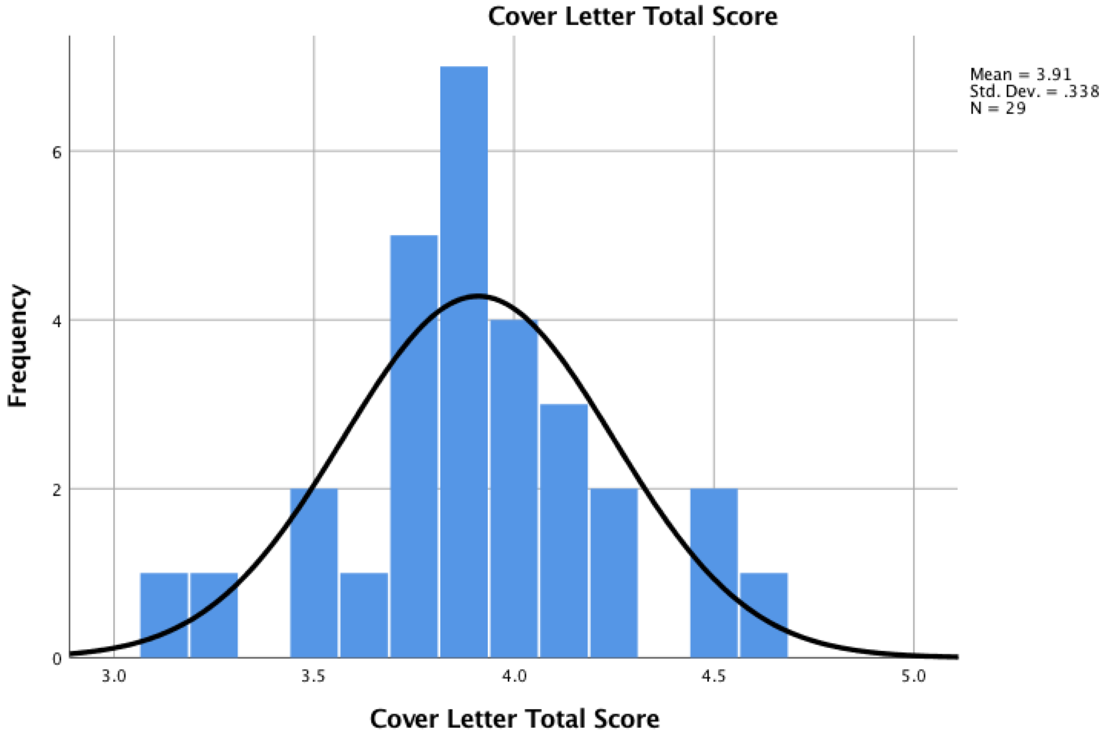


Figure 1 2015: Distribution of scores for cover letter

In 2016 the cover letter and personal statement were not used and instead applicants were instructed to submit written answers to four questions that asked them to address professional issues (Appendix G). The questions covered the topics of medical error, clinical handover, crucial conversations, and advice to junior staff. Each application question was marked by the same assessor out of a maximum of 10 marks. The distribution of scores for the cover letter in 2015 and the total score for all four application questions in 2016 is similar, however there was a substantial difference in the weighting of the two tools. The total contribution of the reflective written component of the application in 2015 was 12.6% for the combined cover letter and personal statement, compared to 27.4% in 2016 for the written application questions. Again, this appears to have been a consequence of the revised scoring system.

Referee Reports

The way referee reports were used differed in each year. Blank proforma reports were emailed to the referees nominated by each applicant (Appendix B). The referee was instructed to score the applicant across a total of ten items in three domains, *Communication Skills, Clinical Competency, Professional and Personal Conduct*. Referees used a 5-point scale from *exceptional performance* to *average* to *requires substantial development.* Referees were asked 'Would you be prepared to have this applicant work for you again?' Applicants could log into an online application system and view whether their referee had submitted the report but could not review the report itself. It was the responsibility of the applicant to ensure their referees submitted the completed report in time.

Different scoring systems were used in each cohort, but this had little effect on the mean or range of scores obtained. In 2015, each applicant's referee reports were marked by all four assessors and an average score used. Assessors were required to globally rate each referee report as either 'poor, average, good or excellent' which determined a final score out of a maximum of seven marks. No information was provided to define these terms. There was a narrow range of scores for applicants which was not useful to differentiate between applicants. In 2016, referee reports were marked *by an individual assessor* using different marking criteria to 2015. *Each* of the ten items were scored across a five point rating scale. Each referee report contributed up to 20 marks to the final selection score. Whether a numerical scoring system was used (2016) or scoring relied on an assessor's global impression of the referee report (2015), a similar range and distribution of scores was reported. The change to the total referee report score increasing to 40 marks meant that it became the most influential tool (along with written application questions) on final scores in 2016.

Selection Centre

There were five selection centre assessments common to both cohorts: Clinical Task, Multiple Choice Questions (MCQs), written responses to clinical scenarios, an admin score and an interview. In 2015 a Communication task was also used. Each item was scored and the total score (selection centre items plus shortlisting items) was used to rank applicants, as described below.

Clinical Task

Applicants were given instructions to complete a clinical task. In 2015 applicants were asked to complete a medication chart to prescribe antibiotics and paracetamol for a patient featured in a clinical vignette. The task required the applicant to correctly document the patient's allergies and weight, the correct dose, route, frequency, and timing of medication. This task was marked by a single assessor and contributed 7.4% to the final score. In 2016 applicants were presented with a scenario and then instructed to complete a pathology form for the patient to order a range of blood tests. The clinical task was scored out of five in 2016 and contributed 3.4% in 2016 to the final selection score. There were a range of scores in 2015 (Figure 2), however in 2016, 97% of the cohort scored full marks (five) for this task with one applicant scoring four out of five (Figure 3).

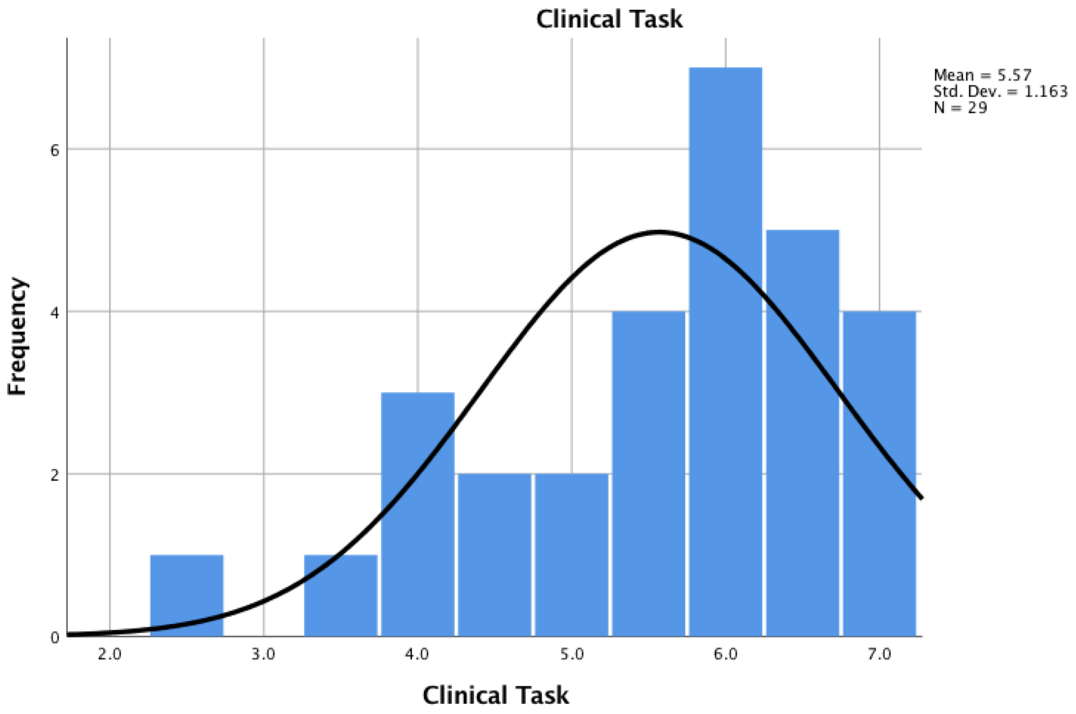


Figure 2: 2015: Distribution of scores for clinical task

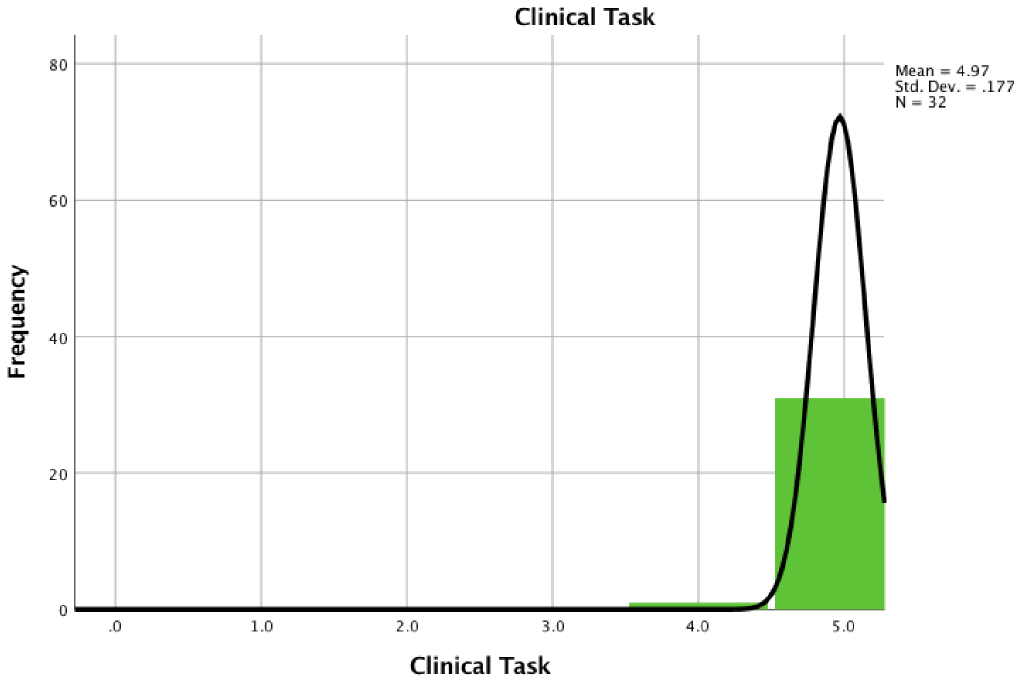


Figure 3: 2016: Distribution of scores for clinical task

Multiple Choice Questions

Seven Multiple Choice questions (MCQs) were asked in each cohort. In 2015, these covered topics in resuscitation, fluid management, management of seizures and respiratory illness, and interpretation of blood gas results. In 2016 these covered neurological assessment, heart disease, consent, resuscitation, fluid management, and differential diagnosis in a case of vomiting. There was a similar range of scores in both cohorts.

Written Response Questions

Applicants were required to provide written responses to two clinical scenarios. These items were designed to assess both clinical knowledge and the applicant's response to dealing with conflict within the clinical team. In 2015, one scenario dealt with the management of a patient who had a complication following a surgical procedure. The other scenario required explaining how they would respond when they needed their registrar's assistance to manage a seizure, but the assigned doctor was refusing to attend. In 2016, one scenario dealt with the clinical management of a patient presenting to the emergency department with fever and abdominal pain. The other scenario asked the applicant to write about their response to a registrar who had given them a patient management plan that they did not feel comfortable with. The mean percentage scores were similar in both cohorts (62% and 61% in 2015 and 2016 respectively).

Administrative Score

Applicants received a score of up to a maximum three marks based on their punctuality, attitude, diligence to completing required paperwork and communication with selection centre administrative staff. Administrative staff allocated this score, and no marking guide was provided. Applicants were not aware this core was being allocated. The selection coordinators included this item to identify objectionable behaviours in applicants (such as had been reported by administrative staff in previous years) and use this information to inform selection decisions.

> …*it's an attempt at creating the opportunity for unmasking undesirable behaviours…And if that score was the thing that made the difference between somebody in and somebody out, I think it's valid. (Interviewee 2, p 27)*

In 2016 it appears as though this item achieved its purpose despite only contributing 2% to the final selection score. An applicant who scored 1.5 on the admin score in 2016 was ranked 15th and was the first person to miss out on a training position. In this case, a score of 3, like the majority of the cohort, would have placed them in the top 14 applicants and they would have been selected.

Interview

Interviews have good face validity with applicants and assessors and it is hard to imagine a selection process without them. A face to face panel interview, with four interviewers was used in both cohorts. Each interviewer scored the applicant out of seven for each question and an average score was used in calculating the total selection score. Interview questions and marking criteria are given in Appendices 3 and 5. The distribution of scores for the interview was broad in both cohorts.

In 2015, a global interview score was used that reflected the interviewer's overall impression of the applicant. The selection coordinators considered this a separate selection tool and there were no marking criteria used. The maximum score available was five, however documented scores ranged from 2.4 to 5.3. This indicates an error in scoring or documentation that was not detected until this analysis. The global score was not felt to be useful and was removed from the process for the 2016 cohort.

Reliability and validity are the two most common measures of a selection tool. Whilst there are no outcomes against which the predictive validity of the interview can be assessed in this case study, intra-class correlations were calculated to measure assessor agreement when rating the same applicant during the interview. Such consistency between raters is viewed as important for the quality and fairness of the assessment. In this case, the correlations for interview questions marked by four assessors ranged from 0.39 to 0.92 (Table 2). As correlations above 0.8 are considered appropriate for high stakes assessment (Downing, 2004), this indicates that for some questions assessors were consistent in their views on the same applicant and in others their opinions were divergent.

Table 2

Interviewer inter-rater reliability measured using intra-class correlation

Item	**Intra-class Correlation***	**Confidence Intervals**
2015		
Interview Question 1	0.89	0.79-0.94
Interview Question 2	0.76	0.57-0.88
Interview Question 3	0.57	0.28-0.77
Interview Question 4	0.39	-0.04-0.68
Global Score	0.75	0.55-0.87
2016		
Interview Question 1	0.92	0.86-0.96
Interview Question 2	0.77	0.59-0.88
Interview Question 3	0.65	0.40-0.81

* Recommended correlation for high stakes selection is 0.80 (Downing, 2004)

Communication task (2015 only)

The communication task used only in 2015 was intended to assess the applicant's ability to perform clinical handover. The applicant was presented with a one-page typed document that had details about an unwell patient with diabetic ketoacidosis.

The task was to read the information and then in a role play, provide clinical handover to an assessor who was playing the role of the doctor taking over the patient's care on the next shift. The assessor was looking for structured, concise communication of the relevant information. This task was planned to contribute 7.4% to the final score. However, there were logistical issues in implementing this assessment. The selection centre was held over two days. Due to staff illness, the assessor changed after the first day. The selection coordinators identified discrepancy in scoring between the two assessors with one assessor giving consistently lower scores (the individual scores for each assessor were not available to the researcher). The distribution of scores is shown in Figure 4. The coordinators decided not to include these scores in the final selection score. The selection coordinators reported that excluding this task from the final selection score did not alter which applicants were selected.

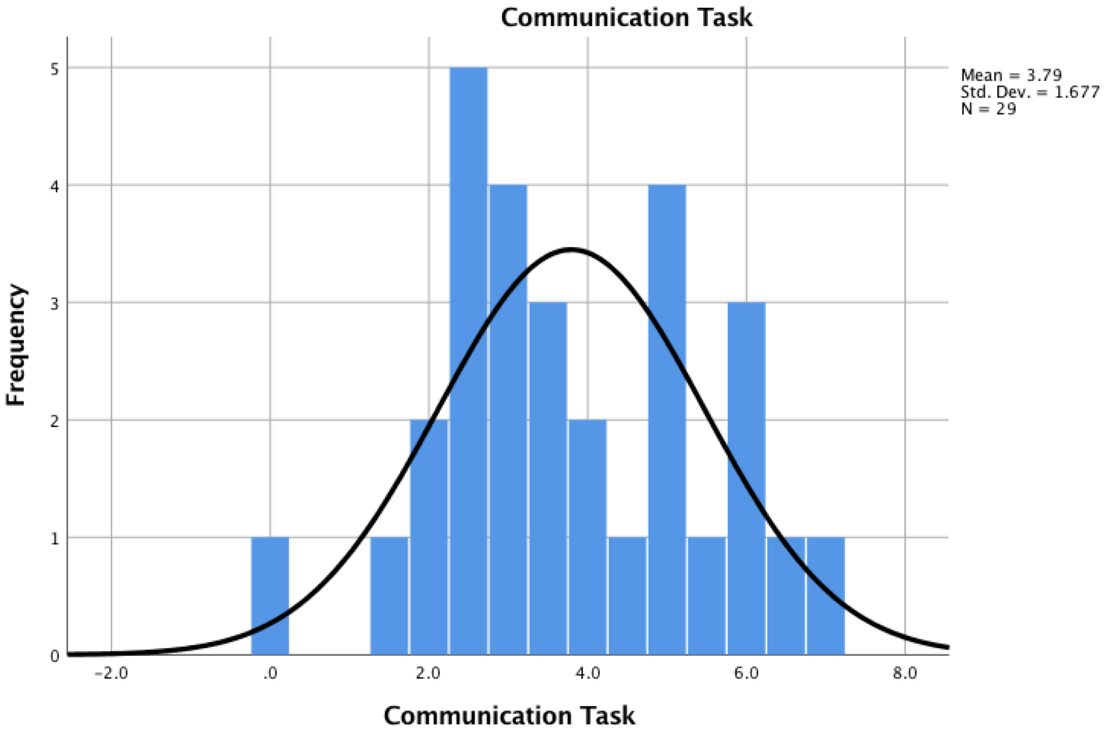


Figure 4: 2015: Distribution of scores for communication task

Making Selection Decisions

This section reviews how information within each application year was used to make selection decisions. Each selection tool assessed applicant performance (usually against set criteria) and provided a score as a measure of their performance. Scores for all selection tools were summed together to reach a final selection score for each applicant. This process reduced all information gathered on applicants to a final summed score. This allowed applicants to be ranked and training positions to be offered in order of the applicants' ranking. The training positions were offered to the top 15 and 14 applicants in 2015 and 2016 respectively.

Given the concerns about justifying selection decision based on scores, the cut-off points for selection (or not) were examined. There was a cluster of applicants with scores around the cut-off points in both cohorts. In 2015 the last applicant selected had a score of 62.6 and in 2016 the last applicant selected scored 105.3, as shown below in Table 3. The distribution of scores and the selection cut-off point for both cohorts are shown in Figures 5 and 6, showing a significant number of applicants grouped around these cut-off scores.

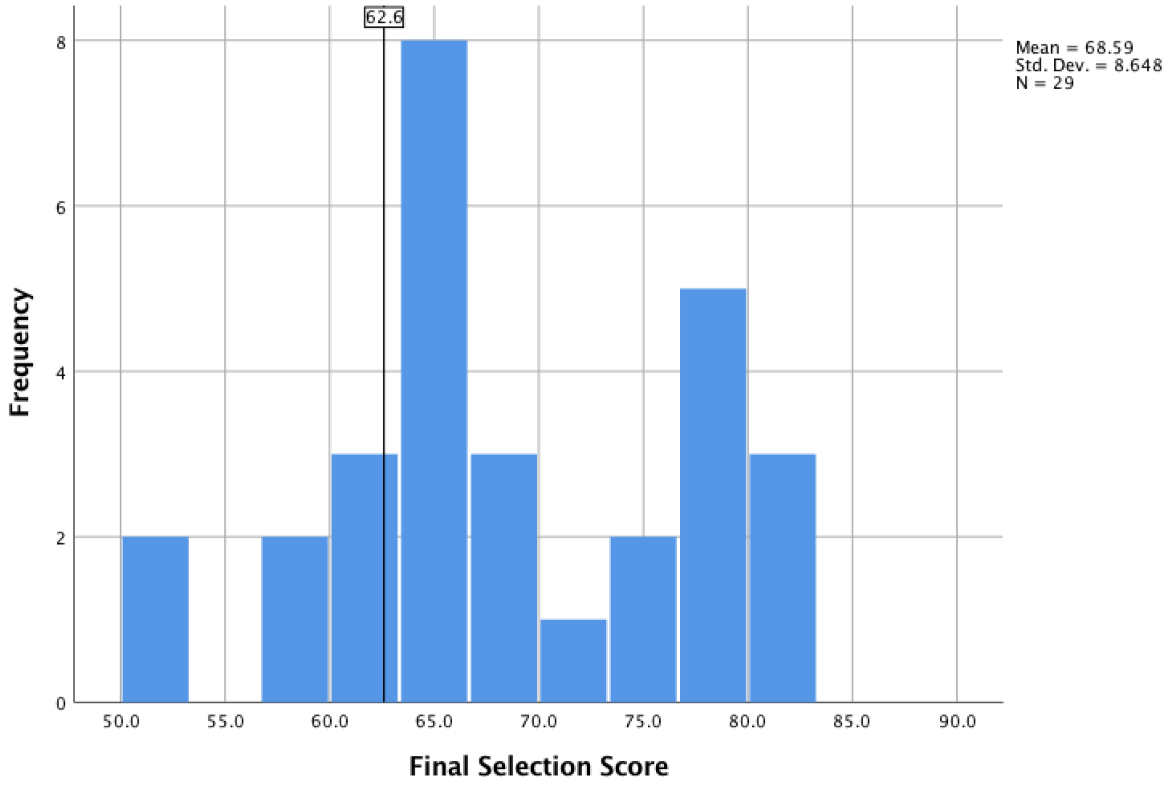


Figure 5: 2015: Distribution of final selection scores
(Selection cut point score of 62.6 is shown)

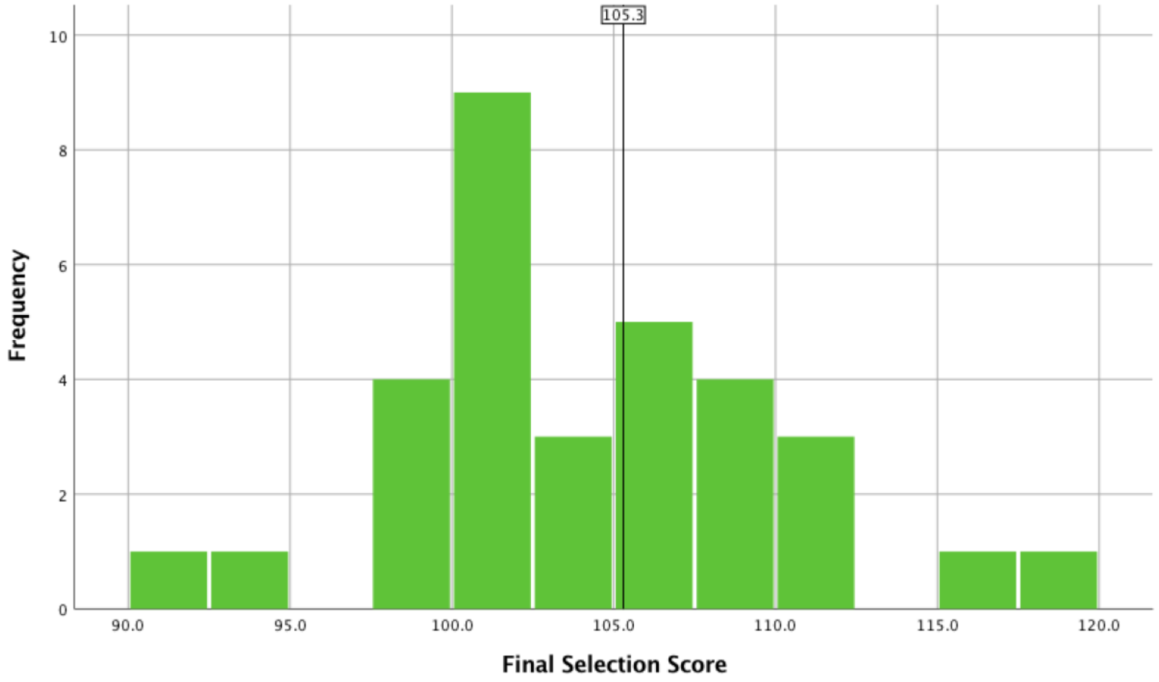


Figure 6: 2016: Distribution of final selection scores
(Selection cut point score of 105.3 is shown)

Table 3

Final selection scores

(Successful Applicant Scores are Shaded)

2015		2016	
Rank	**Score**	**Rank**	**Score**
1	76.3	1	118.8
2	75.4	2	115.8
3	75.1	3	111.3
4	75.1	4	111.3
5	74.4	5	110.1
6	74.1	6	108.8
7	73.8	7	108.5
8	73.4	8	108.3
9	72.5	9	107.8
10	69.6	10	107.0
11	69.0	11	106.5
12	65.9	12	106.5
13	64.8	13	106.3
14	63.1	14	105.3
15	62.6	15	103.1
16	62.3	16	102.8
17	62.3	17	102.5
18	62.0	18	102.4
19	61.9	19	102.4
20	61.0	20	102.3
21	59.6	21	102.3
22	59.0	22	102.0
23	58.8	23	102.0
24	58.8	24	101.6
25	57.6	25	100.5
26	57.1	26	100.4
27	54.9	27	99.5
28	51.5	28	99.0
29	47.5	29	98.6
		30	98.5
		31	93.9
		32	92.0

With such a small number of marks separating the successful and unsuccessful applicants, small changes in weighting can have major effects on each applicant's rank. Exploring how different combinations of selection tools affected selection outcome forms part of this evaluation.

Alternative Selection Models

Modelling different combinations of tools can help in understanding the contribution of these tools and assist coordinators in planning future selection processes. Alternative combinations of tools were compared to the entire suite of tools used in in each cohort. The Kappa statistic was used to measure agreement between the selection process used by SWSN and the alternative models. The results are summarised in Table 4 and 5. These tables show five alternative combinations of selection tools and the contribution these tools made to the original final selection score is listed as a percentage.

The alternative modelling shows that even if different combinations of scores are used, or some scores are left out altogether, most of the same people are selected for training. The kappa coefficient provides a measure of agreement between binary outcomes, and there was at least moderate agreement occurring between all alternative combinations of tools and the SWSN process. Of note, there is a pattern seen that alternatives with a high percentage weighting contribution to the final score, tend to result in a high number of the same applicants still being selected for training (e.g. *Alternatives 3,4 and 5* in 2015). Opposing this trend are the results for *Alternative 2* – making selection decisions on shortlisting items only. In 2015, shortlisting tools only contributed 17.8% to the final selection score yet using them alone would still select 80% of the same applicants. In 2016, when shortlisting tools contributed a much larger 68.5%, there was the lowest level of agreement with the original process. This suggests weighting of selection tools alone is not the only factor in determining selection outcome.

Table 4

2015: Comparison of alternative selection methods to current SWSN process

Alternative Combination of Tools		Contribution of tools to original final score (%)	Applicants Still Selected (%)	Applicants Not Selected (%)	Kappa	Standard error	Confidence Intervals (95%)
1	Traditional Method: (CV, cover letter, interview, referee report)	47.3	80	20	0.586	0.151	0.29-0.86
2	Shortlisting process only: Pre-selection centre items (CV, cover letter, and personal statement)	17.8	80	20	0.586	0.151	0.25-0.86
3	Selection Centre Items only	74.9	100	0	1	0	1.0 -1.0
4	SWSN Process without referee scores	92.7	100	0	1	0	1.0-1.0
5	SWSN Process without interview scores	70.5	93	7	0.862	0.094	0.65 – 1.0

Level of agreement: Poor if κ <0.00, Slight if 0.00 < κ < 0.2, Fair if 0.21 < κ < 0.4, Moderate if 0.41 < κ < 0.6, Substantial if 0.61 < κ < 0.8, Almost perfect if κ > 0.8, Perfect κ = 1

Table 5

2016: Comparison of alternative selection methods to current SWSN process

Alternative Combination of Tools		**Contribution of tools to original final score (%)**	**Applicants Still Selected (%)**	**Applicants Not Selected (%)**	**Kappa**	**Standard error**	**Confidence Intervals (95%)**
1	Traditional Method: (CV, interview, referee report)	55.5	79	21	0.619	0.14	0.30-0.87
2	Shortlisting process only: Pre-selection centre items (CV, 4 written application questions, referee reports)	68.5	71	29	0.492	0.155	0.19-0.80
3	Selection Centre Items only	31.4	71	29	0.613	0.141	0.33-0.88
4	SWSN Process without referee scores	72.6	79	21	0.619	0.14	0.30-0.87
5	SWSN Process without interview scores	85.6	86	14	0.746	0.119	0.47-0.94

Level of agreement: Poor if κ <0.00, Slight if 0.00 < κ < 0.2, Fair if 0.21 < κ < 0.4, Moderate if 0.41 < κ < 0.6, Substantial if 0.61 < κ < 0.8, Almost perfect if κ > 0.8, Perfect κ = 1)

The decision-making model used in the case study meant that low scores in one tool did not preclude an applicant from being selected. There were several instances where an applicant scored poorly in a particular tool and was still successfully selected (Table 6). Notable examples are successful applicants scoring 2.5 for the clinical task in 2015, and a score of 11.3 for interview in 2016. There was no mechanism to review individual tool scores when making selection decisions, the summed total of tools scores determined the final rank list.

Table 6

Lowest scores by tool for successful applicants

	2015 n = 29			**2016 n = 32**		
Selection Tool	**Max Score**	**Range**	**Lowest Score by a Successful Applicant**	**Max Score**	**Range**	**Lowest Score by a Successful Applicant**
Curriculum Vitae	5	3.3-4.9	3.3	20	11-17	13
Personal Statement	7	3.4-5.9	3.4	-	-	-
Cover Letter	5	3.1-4.6	3.1	-	-	-
Application Questions	-	-	-	40	19-30.5	23
Referee Reports	7	4.5-7	4.5	40	27.5-37	30.5
Clinical Task	7	2.5-7	2.5	5	4-5	4
Multiple Choice Questions	7	3-7	5	7	4-7	5
Written Response Questions	14	2-13	4	10	4-8	5
Admin Score	3	2.5-3	2.5	3	1.5-3	2
Interview	28	13.4-24	16.8	21	9-19.75	11.3
Global Interview Score	5	2.4-5.3	3	-	-	-
Communication Task	7	0-7	2.5	-	-	-

Social Validity

The social validity refers to the fairness of a selection process as perceived by the stakeholders. As part of the evaluation, views were sought from short listed applicants using an exit survey given to those who attended the selection centre in 2016 (a response rate of 81%). The results (Table 7) show that applicants thought the overall process was appropriate, fair and provided the opportunity to demonstrate their skills and abilities. The traditional methods of CV, referee reports and interview were endorsed as being appropriate and fair. Selection centre tasks (MCQ, written responses and pathology form) had lower levels of agreement but were all still rated positively by applicants. In general, the selection process used by the SWSN was well received by the applicants. Free text comments on the process were generally positive, a sample of comments presented in Figure 7 illustrate the general trend in applicants approving the selection centre approach.

Table 7

2016: Summary of applicant exit survey data

	Component of Selection Centre % Agreement (Agree or Strongly Agree)		
Selection Tool	**Task was Appropriate**	**Opportunity to demonstrate skills**	**Task was Fair**
Curriculum Vitae	100	96	100
Application Questions	92.3	95.8	100
Referee Reports	96	96	96
Clinical Task	72	88	100
Multiple Choice Questions	88	83.3	90.9
Written Response Questions	76	76	87.5
Interview	100	95.7	96

Very refreshing method. Takes pressure off just an interview. Good balance of required clinical knowledge and professional skills.

Far better process than simply having an interview and CV for application.

MCQ was a good way to screen for basic specialty knowledge

Short answer questions more objective than cover letter

Figure 7: 2016: Quotes from applicant exit survey

Chapter Summary

This case study has provided an opportunity to examine the data collected in a vocational medical selection process. The key findings in the chapter are concerned with how observations and assessments of applicants were converted to scores, and how these scores were used to make selection decisions. A range of selection tools were used, with modifications made between cohorts to enhance perceived fairness and for logistical reasons. These changes made little difference to the applicants' performance in terms of the mean and range of scores (with the exception of the clinical task). However, changes to scoring had a significant effect on tool weighting. The decision-making process required tool scores to be summed to produce a rank list, with several applicants receiving scores close to the numerical cut point. When different combinations of tools were modelled, most of the same applicants would be selected with the weighting of tools playing an influential role. Social validity of a selection process is important, and the exit survey indicated that applicant perceptions of fairness were favourable.

The findings of the group interview (Chapter 4) and the case study evaluation (Chapter 5) and brought together for discussion in Chapter 6.

6. DISCUSSION

This chapter is split into two parts. First, the findings of the case study from Chapters 4 and 5 are discussed. The second part of this discussion chapter explores whether the principles of Programmatic Assessment could be used as a framework for selection.

6A. DISCUSSION OF THE CASE STUDY

The case study provides an insight into the challenges of design, logistics and decision-making in vocational medical selection. The group interview revealed that the selection coordinators were motivated to have a fair process that selected the best trainees. They acknowledged that assessing whether they were selecting the right trainees (predictive validity of their process) was difficult. Like many case reports in the literature, they were concerned about face validity and perception of fairness throughout the process. Pragmatism played a role in shaping the process, as they needed to select a limited number of trainees from a large pool of applicants. The key findings of the evaluation concern the way information obtained from selection tools is quantified and used in decision-making frameworks.

The Quantification of Selection Tool Data

The selection coordinators operated in a positivist paradigm and this influenced the way tools were used to make selection decisions. A positivist view of selection holds that it is possible to identify the best applicants and that the quality of those applicants is objectively measurable. This was evident in the preference for using numerical scoring rather than gestalt impressions, or their own judgement based on previous experience. For example, global assessments of shortlisting items in 2015 were seen as prone to bias. The change to using a scoring system for these items in 2016 was viewed as an improvement in objectivity. In the same way, using the summed score of all selection tools to produce a rank list demonstrates a trust in the numbers; that a larger total score truly represents a higher quality applicant. The reliance on numerical measurement has several consequences not unique to this case study.

The quantification of qualitative observations means potentially valuable information is hidden from selectors. Each tool collected information that could provide some insight into the applicant's suitability to pursue vocational training. This included, the content of their written responses that demonstrated knowledge of clinical management and professional issues, incorrect answers on the MCQ which highlighted knowledge gaps, and thoughtful interview responses that the marking schema may not record. Even if such information was essential for determining suitability for training (or for evidence that applicants were unsuitable for training) it was only the summed score that was used to make selection decisions. Information provided in applicant responses not revisited by the selection panel.

Referee reports provide an example of where potentially useful information about applicants is hidden from the decision-making process because of the scoring system used. It has been suggested that referee reports that provide negative information are the most useful (Patterson, Ferguson, et al., 2013). A review process was used in 2016 to consider comments made by referees. Comments made by the referee were only considered if they were inconsistent with the scoring. The referee report score could not reflect such negative comments however, as the scoring system was based purely on the 5-point scale used to assess items in each domain. Any insights a referee may have into an applicant's ability were lost when the referee report was converted to a numerical score. Even if the referee stated they had concerns about working with the applicant in the future, the score was unable to reflect this.

Despite the selection coordinators universally converting all information into a score, they clearly valued the more detailed information obtained in the tools. They reported that in 2015 the cover letters all tended to contain the same generic content and did not provide useful information about the applicant.

> *We've looked at it and got an impression of what was (sic) the most discriminating exercises, and it looked like some were far more discriminating than others; that's why the cover letter was dropped. Basically most cover letters are just cut and paste from the CV, so the same thing. (Interviewee 2, p15)*

In 2016 four written responses to application questions were used instead of the cover letter to improve the quality of information provided by applicants. The translation of this information into scores, meant valuable content about applicant knowledge was lost and not able to be considered in selection decisions.

Interestingly, there is a paradox about the selection coordinators' views on the value of human judgement and narrative information. They were concerned about using too much 'subjective' human judgement when assessing applicants and preferred the certainty of applying a numerical score. There was trust in the numbers, and final scores and rank lists were not questioned or reviewed. However, when it came to assessing the merits of selection tools, they were content to use opinion and individual impressions to design and later modify the selection process, and did not use any psychometric data from the tools. They modified the number of assessors for each tool and trialled different scoring rubrics. Increasing the number of tools was a strategy to provide a greater spread of scores to facilitate decision-making processes, rather than strategic use of tools to assess specific attributes. While the published case reports described in the literature review sought to improve their selection processes by optimising the psychometric properties of the tools used, the selection coordinators in this case study made modifications to expedite perceived fairness and efficacy.

Reductionist Decision-making Framework

Another consequence of the positivist view of selection is that it begets a reductionist approach to decision-making. The framework used by selection coordinators left them with continuous data (the final selection score) which they had to use to make a dichotomous decision (selection or non-selection). A numerical cut point was needed to separate the last to be selected and the first to miss out. In both cohorts there was a significant number of applicants clustered around the cut-off point and selection coordinators were anxious that the scores should provide clear delineation of who was selected and who was not. With this approach the weighting given to each tool becomes an important consideration.

Weighting of Selection Tools

In this reductionist decision-making model, the weighting of individual selection tools was found to be the most influential factor in determining selection, rather than the method of measurement or the content of the tool. In 2015, the most influential tool on selection outcome was the interview at 29.5% of the total final score. With changes made to the process in 2016 this was halved. In 2016, referee reports and written response questions contributed the largest scores at 27.4% each. In fact, 68.5% of the total score in 2016 came from shortlisting items submitted before the selectors met the applicants. These weightings had a substantial impact on who was selected.

The influence of weighting was evident in the modelling of alternative combinations on selection tools in Chapter 5. *Alternative 3*, using selection centre items only, provides a useful example to discuss this. In this model, all decisions about selecting into the training program are based on performance at the selection centre. Scores for shortlisting tools are not considered in this model. In the 2015 cohort, using selection centre tools only would result in perfect agreement with the current process that uses all tools. That is, the same 15 applicants would have been selected using the selection centre scores alone, as were selected using all scores. In 2016 however, only 10 of the same 14 (71%) applicants selected would still have been selected if only selection centre scores were used. This difference is likely related to the higher weighting given to the selection centre in 2015 (74.9%) compared with 2016 (31.4%). A similar situation occurs when using *Alternative 4*, the model of making selection decisions without the use of referee scores. In 2015, removing referee reports from the final selection score made no difference to selection outcome. In 2016, when referee reports had an inflated weighting, removing them resulted in three of the selected applicants missing out on selection. Given the influence of tool score weighting on final summed scores coordinators need to be cognisant of this when configuring scoring and decision-making processes.

Interestingly, going against the argument that weighting is a major factor is *Alternative 2* in 2016; using only shortlisting items to make selection decisions. We would expect a high level of agreement between this model and the existing SWSN process given that these shortlisting items contributed 68.5% to the final selection score. However, this was not the case, as this model had the lowest level of agreement. This provides some justification for the resource intensive selection centre. Given the low level of agreement it is probable that the selection centre items in 2016 assessed something different to the shortlisting items in the same cohort.

The weighting of tools is not a sufficient explanation for all results of the alternative modelling. Using a different set of selection tools and selecting the same applicants may be explained if the tools are actually assessing the same constructs. Although the tools were not mapped to specific constructs in the SWSN process, it is possible that they are inadvertently measuring the same applicant characteristics. This is plausible considering the content of the tools chosen. For example, the MCQs, written scenarios and the clinical task all test aspects of clinical knowledge. For other tools, like answers given in written application questions and performance at interview, the links are not immediately obvious. Factor analysis can be used to determine whether particular tools are potentially examining the same construct (Cook & Beckman, 2006). This was considered in this study, but the cohort size was too small for meaningful use of this analysis. Deliberate mapping of constructs tools in the design of the selection process would help coordinators to understand how performance in difference tools is linked.

The rationale for different weightings was unclear in this case study. The weighting given to each tool was a consequence of the scoring system for each tool, rather than a deliberate decision to weight some selection tools more heavily than others. A consequence of expanding the range of scores for shortlisting items (CV, written responses in 2015 and referee reports in 2016) was the substantial weighting given to these tools. This weighting was out of proportion given there is little evidence to support their reliability, validity or ability to predict future job performance.

Compensatory Decision-Making Framework

The influence of each tool in this case study was determined by its weighting because of the compensatory decision-making framework used. Poor performance in one tool could be recovered by good performance in another tool with a heavier weighting. As was shown in

Chapter 5, Table 6, there are several examples where this occurred. The clinical task in 2015 required applicants to demonstrate ability in a basic, essential skill for a trainee: prescribing, and there was a good spread of scores. One might assume that this is a useful task to separate those applicants not suited to training. However, on review of final selections, it appears a low scoring applicant was still successful in being selected because of higher scores in other tools. Likewise, in the written responses to clinical scenarios, applicants with low scores for questions concerning clinical knowledge and professional behaviour were still able to be selected. This is disconcerting in a situation where unsafe practice or lack of basic medical knowledge can be compensated for by a well written answer in another tool. These examples show how the use of summative compensatory scoring can undermine the use of tools designed to measure important constructs.

Interviews have good face validity with applicants and selection coordinators, so it is interesting to consider their value to the selection process in a compensatory model. A review of the data reveals that if only the interview were used for selection, around 80% of the successful applicants in each cohort would have been successfully selected. This means that through using the other items in the selection centre, around one fifth of successful applicants were not amongst the top performers at interview and were able to 'earn' their selection through their performance in other tools. It is also true then that performance at interview correlated with success in the selection process as a whole. There are a number of interpretations for this. One is that performance at the interview in this case was indicative of suitability for training. However, there are no predictive validity studies to prove that indeed those selected for training in the SWSN were actually suitable. Another interpretation is that the correlation between interview performance and selection outcome is a function of the weighting of the tools. The interview contributed 29.5% of marks to the final score in 2015 and 14.4% in 2016. This means that scoring well at interview will benefit an applicant's final selection score considerably compared to other selection tools, and therefore increase the likelihood of selection. However, it was still possible to be selected and perform poorly in the interview. In 2015 an applicant ranked 26th of 29 by interview score was still selected in 2016 an applicant ranked 28th of 32 by interview was selected.

Interviews are resource intensive and in large scale selection processes training is required for interviewers. If interviews are measuring constructs that can be assessed using other tools that are easier to administer, then coordinators may consider removing interviews from the process. This decision would need to balance the feasibility of interviews against the acceptability to stakeholders of removing them from the process. Interviews may also have a role in selecting out unsuitable applicants, but this is difficult when a compensatory decision-making framework is used.

Social Validity

Applicant perception of the appropriateness (political validity) and fairness (social validity) was assessed through an exit survey. Overall the results were favourable (p90). This should be reassuring for selection coordinators as the changes they made to the selection processes were in part motivated by a desire to improve the fairness overall. Also, selection processes with low levels of stakeholder acceptability are more likely to be challenged with appeals (Anderson, 2011). These results should be interpreted with some caution and caveats. The survey results are only an indication of perceived fairness with regard to the content and operation of the selection tools. Applicants were not aware of how selection tool information was quantified and then used (or not) to make selection decisions. Stakeholder acceptability of this aspect of selection is unknown. The evaluation in the case study was also limited to two stakeholder groups, applicants and selection centre coordinators. The views of patients, nursing staff, supervisors of trainees and fellows of the college or important and often missing from evaluations published in the literature (Kelly et al., 2018). There is the opportunity to explore the political validity of the SSWN process further.

The initial research question in this thesis was to identify an effective process for vocational medical selection. From the case study and literature review we know that selection is challenging and the factors that lead to it being effective are complex. There are many variables to consider, and a framework is needed to design and organise the different elements of selection. Focussing on the format, content, scoring system and psychometric rigor of selection tools is only one part of the overall selection process. Attention needs to be given to aspects of selection beyond the tools. These elements include the purpose of the selection, the desired attributes for trainees and how selection tools can be used to assess these. A framework is needed that provides direction on how to use the information gathered about applicants to make selection decisions and also to give guidance on evaluation of the process and quality assurance mechanisms. The next section discusses whether an assessment framework can provide the structure and practices to design effective selection processes.

Limitations

This case study provided a useful platform to explore vocational medical selection with some limitations that should be acknowledged. The case study operates in the Australian context and within a state-based selection system. While similar selection systems are known to operate in the UK and New Zealand, there may be some aspects that are less applicable to other contexts with direct entry from medical school to specialty training (e.g. North America) or to selection programs that operate at a national level. Still the principles regarding the proposed decision-making frameworks and the way information from selection tools is combined, remain relevant. This sample population reflects the ‘real world’ situation faced by many selection coordinators of vocational medical programs. There are many sites that must deal with same complexities as the SWSN and design a process to select trainees into their programs.

Data was available for two cohorts of applicants and overall the sample size of applicants was relatively small. This prevented the use of some statistical procedures (factor analysis, discriminant analysis) that may have provided useful information on the discriminatory value of selection tools and the constructs being assessed. The wide confidence intervals for the statistical tests used (kappa coefficient, intra-class correlations) are consistent with the small sample size and are a threat to the validity of any conclusions about significance drawn. It would not be appropriate to combine data to increase the sample size as each cohort used different types of tools, with different content and scoring systems and selection decisions were made based on ranking within the cohort so combining cohorts would not be valid. Despite the small sample size, what is useful from this analysis is the understandings gained about influence of the tool weightings on selection outcome. The quantitative post hoc evaluation of the section tools performance provides useful information to inform local design decisions.

6B. A FRAMEWORK FOR SELECTION

This second part of this discussion chapter explores whether a Programmatic Assessment model, which emphasises programs of assessment rather than individual methods, could be used as a framework to design and implement selection processes. The case study presented illustrates many of the complexities and challenges of vocational training selection that were identified in the reviewed literature. Many of these same challenges are faced in assessment. The following discussion explores the common features of assessment (in medical education) and selection and highlights the key differences. The Programmatic Assessment model is presented and considered as a possible framework for designing and implementing selection processes.

This project set out to answer the question:

What is an effective process for selecting applicants into vocational medical training programs?

Both the literature review and the case study revealed that the precise ingredients for an effective selection process are unclear. There is an emphasis on picking the particular selection tools with MMIs and SJTs promoted as having ‘good evidence.’ However, the utility of a selection tool is determined by many variables including the content, format, scoring matrix and level of assessor training. In the case study these variables were often modified with the aim to improve fairness and summative efficacy, such as identification of a clear cut-point for selection decisions. In the literature, the aim of manipulating these variables was often to improve reliability of individual tools. Despite a burgeoning literature on individual selection tools, guidance on how these tools should be used together to make selection decisions is lacking. Reductionist approaches to decision-making mean qualitative information is lost when converted to a score. Numerically combining scores means that the weighting of scores for each tool becomes a significant determinant of selection outcome as demonstrated in this case study. In view of the challenges of the designing selection process and effective utilisation of information gathered, there has been a call for new frameworks for selection (Roberts 2017, Patterson 2016, Prideaux 2011).

Assessment and Selection

Many of the principles of assessment in medical education could be applied to selection. These include documenting a plan for a selection process that links desired attributes to assessment methods, clearly defining standards and decision-making processes and using evidence from psychometric studies to inform process design (Prideaux, 2011). At face value, selection and assessment are very similar processes. Figure 1 shows a comparative schema of each. In both cases a group of people (learners in assessment, applicants in selection) undergo an evaluation of their competence using a range of tools. Information from these tools is used to determine an outcome. In assessment this can be a graded outcome (credit, distinction, high distinction etc.). In selection a binary outcome is reached, that is, selected or not selected.

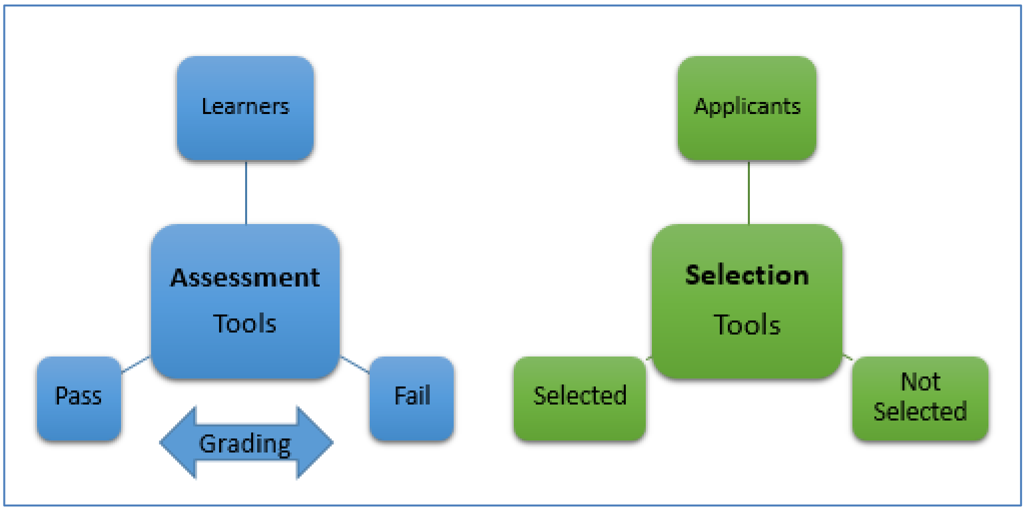


Figure 1: Assessment and selection schematic

However, there are important differences between assessment and selection (Table 1). Both processes seek to determine if learners or applicants have reached a level of competence. Yet selection processes have a number of other roles too. They must identify and 'select out' applicants who may meet criteria for being competent in some areas but are unsuitable for selection due to other issues (e.g. lack of professionalism) (Roberts & Togno, 2011). There is also a role in widening access to training programs for particular groups (Patterson, Cleland, & Cousans, 2017). A fundamental difference is that in assessment all learners assessed as competent will pass, while in selection not every competent applicant can be selected, due to limited positions available. Ultimately unless there are unlimited places available, a selection process must differentiate amongst a group of competent applicants and make a dichotomous decision about selection.

Both systems operate in different paradigms. In assessment an evaluative paradigm is used, with a post hoc review of a learner determining whether skills and knowledge *have* been acquired. The predictivist paradigm in selection needs to establish if an applicant has the potential to be trained in the future. (Patterson, Ferguson, et al., 2013) The institutions in each process have different roles. In assessment, the same institution (University) is responsible for teaching the learner and assessing the outcome of that teaching. The goal is for all learners to achieve an exit level of competence. With selection, the institution responsible for assessing applicants for entry into training (medical specialty college) usually has no responsibility for the education for the applicant up to that point.

Table 1

Comparison of medical educational assessment and selection

	Educational Assessment	**Selection**
Purpose/ Primary Outcome	Determine if learners have achieved a requisite level of competence	Determine if applicants have achieved a requisite level of competence AND Discriminate between applicants who are assessed as competent AND Widening Access AND Select out unsuitable applicants
Paradigm	Evaluation paradigm: determine whether a learner has been trained successfully	Predictivist paradigm: determine whether applicant will be able to be trained in the future
Role of assessing system	The same institution is often responsible for training and assessing learners. System designed to encourage all learners to achieve an exit level of competence	The institution assessing applicants is usually independent from the organisations where the applicant has trained to develop their skills. System designed to assess competence, no role in assisting applicants to achieve competence
Feedback	Learners are able to benchmark performance through a grading system. Feedback can assist with performance improvement.	Binary feedback: selected or not selected.
Stakes	High stakes assessment Potential for remediation Theoretically all students can pass	High stakes assessment Remediation is not offered Not all applicants can be selected

The evolution of selection practices used in vocational medical training has followed a similar course to medical education assessment, albeit with some years delay. Van Der Vleuten's (1996) commentary on the state of assessment in medical education discusses a number of issues that are relevant to selection practices used for entry to vocational medical training today. Rising numbers of students in higher education since the second world war presented logistical issues for assessment practice (van der Vleuten, 1996). The explosion in the number of medical students over the last decade (Medical Deans Australia and New Zealand, 2015) similarly presents challenges for selection for limited vocational medical training places. In medical education, traditional assessment practices had relied on an apprenticeship model, with competence assessments based on the holistic opinions of preceptors and unstandardised tests. Assessments were seen to be too *subjective* which was viewed negatively and criticised as being open to bias (Hodges, 2013). This approach is also seen in selection with clinical specialists undertaking unstructured panel interviews with rudimentary scoring systems (Goodwin et al., 2014). In both areas this led to efforts to formalise processes and focus on quantitative measures and the psychometric approaches to assessment.

Part of the formalisation of both assessment and selection has been the matter of defining competence. In medical education, competence was considered to be a collection of attributes (Schmidt, Norman, & Boshuizen, 1990) and that each attribute could be measured by a different method (e.g. MCQs to assess knowledge, observed structured clinical examination (OSCE) to assess clinical examination skills) (van der Vleuten, 1996). This path has been followed in vocational medical selection particularly in the UK, Australia and New Zealand, with the adoption of the principles of competency based medical education (Frank et al., 2010; Patterson, Tavabie, et al., 2013). The key characteristics required for the training program are identified through job analysis and consultation with stakeholders within the specialty (Gale et al., 2010; Patterson et al., 2008). These attributes are often separated into academic and non-academic (e.g. integrity, empathy, team work) domains and then mapped to selection tools (Roberts et al., 2017) (Patterson, Ferguson, et al., 2013).

A positivist view has emerged in both fields. Because it was perceived that competence could be measured, the research agenda became a search for the best measurement method for different attributes (Schuwirth & Van der Vleuten, 2011). Seeking to optimise tools to measure individual traits and quantify them with a score has been described as part of the rise of the psychometric discourse in assessment (Schuwirth and van der Vleuten 2006). Educational research focussed on the reliability of assessment tools and finding the optimal number, type and duration of testing items (Hodges, 2013; Newble, Baxter, & Elmslie, 1979; van der Vleuten, 2016). The selection literature too is dominated by studies reporting on the psychometric properties of measurement tools (Bandiera & Regehr, 2003; Goodwin et al., 2014; Goodyear et al., 2007). In selection, MMIs have been extensively researched with different variations in number of assessors, stations and duration explored to augment reliability (Knorr & Hissbach, 2014) Several authors have cautioned the emphasis on the statistical properties of tools (Hodges, 2013; Kuper, Reeves, Albert, & Hodges, 2007; Schuwirth & van der Vleuten, 2006).

Critique of the Psychometric Approach

Psychometric approaches to assessment have received much attention and critique in the literature. Concerns include the quantification of qualitative data, the reductionist approach to combining data from different measurement tools and the conflicts between reliability and validity when designing measurement tools, as discussed below. Similar issues were seen in the SWSN case study.

Assessment and selection systems both highlight the shortcomings of measurement methods that attempt to quantify human phenomena. The challenge of how to measure, rate and compare human behaviour has long been debated (Schoenherr & Hamstra, 2016). Evaluation of the SWSN case study found that specific constructs were not overtly mapped to the measurement tools. However, when looking at the content of the tools one can conject about the constructs that are potentially being tested. For example, in 2016, written responses in the application would permit assessors to form a picture of the applicants' insight and ability to self-reflect, their conscientiousness, and some information about their communication style (or their own perceptions of their intended communication style as this was a written response). Thus, responses could potentially contribute to judging an applicant's suitability for training. However, this information was appraised using a scoring rubric and converted to a number. This assumed that an applicant with 8 out of 10 would be more conscientious than an applicant who scored 7. Not only can the validity of this be questioned, but rich material contained within the responses was lost to selection decisions. Such 'waste of information' is also described in the use of MCQs for assessment (Schuwirth & van der Vleuten, 2006). Much can be discovered about a learner from their responses to MCQs; which areas they are knowledgeable in, where they need to improve, and what misunderstandings they may have about key concepts. However, this information is not accessible when the result is reported as a percentage of correct answers. This loss is compounded when scores from different tools are combined.

When assessment and selection systems sum the information from multiple constructs (or tools), the value of those individual measures is weakened. In medical education, the criticism of this approach is that learners are able to compensate for lack of knowledge in one area by better knowledge in another (Hodges, 2013). This can have significant ramifications if learners fail at patient management but excel in communication and their scores average out to allow them to progress to qualification. The stakes are likewise high in selection. In the SWSN case study for example, we saw that an applicant could fail a task where they were asked to prescribe medication but compensate for this by scoring well in their CV that self-reports their achievements, and still be selected for training. To illustrate the illogicality of this approach, Schuwirth and van der Vleuten make the comparison to clinical medicine (Schuwirth & van der Vleuten, 2006; van der Vleuten & Schuwirth, 2005). They describe this approach as akin to making a diagnosis for a patient by adding their blood pressure and their sodium level (which is nonsensical). Instead, doctors consider the meaning behind those two measurements beyond the numerical value and combine this with other information from the patient's records to reach a conclusion about their state of health. A similar approach in assessment is advocated (Schuwirth & van der Vleuten, 2006).

A further concern about the psychometric discourse in assessment refers to the conflict between optimising reliability and validity. Reliability is a highly regarded property of assessment and selection methods, and is viewed as a proxy for the quality of a test and also an indication of fairness (Hodges, 2013). The standardisation of tests and the training and calibration of assessors in selection, have been driven by the quest to optimise reliability (Bandiera & Regehr, 2003; Knorr & Hissbach, 2014). The scenarios presented, and responses expected become tightly regulated in order to maximise the reliability of the measurement tools. This comes at a cost of construct validity. Seen through a constructivist paradigm: applicants, assessors, patients, hospitals and employers are not homogenised and a range of responses will be appropriate in different situations depending on context (Kuper et al., 2007). Designing assessment tools that allow for variance in response will inevitably weaken reliability but enhance authenticity to real life situations (Hodges, 2013). In the case study, there were changes made to avoid assessor variance, for example changing to a single assessor for written responses in the application. Consistency of assessment measurements was prioritised over using a diversity of opinions to assess applicants. The danger with the emphasis on reliability is that it is pursued at the cost of validity, leading to tools that do not provide meaningful data on which to make decisions.

Paradigm Change: Programmatic Assessment

Resolving these problems in assessment requires a change in paradigm. Programmatic Assessment is an approach to the design of assessment programs that seeks to address the concerns of the traditional assessment model that emphasises psychometrics. The term *Programmatic Assessment* as used in this thesis refers to the model (or framework) proposed by van der Vleuten and Schuwirth, first formally introduced in 2005 (van der Vleuten & Schuwirth, 2005) and then further described and expanded over the next decade and beyond (Schuwirth & Van der Vleuten, 2011; van der Vleuten, Schuwirth, Driessen, Govaerts, & Heeneman, 2014). Programmatic Assessment is a different way of looking at measurement and the use of information. In essence, it is a set of design principles that involves a systematic collection of data about a learner. There is the deliberate choice of tools based on their purpose and content, both aligned to curriculum outcomes (or competencies or domains). One tool can assess a number of domains and one domain can be assessed by a number of tools. The focus is shifted from the individual assessment tools and their psychometric properties, to the overall program of assessment, with a combination of different data collection methods. Thus assessment events and decision-making events are separated. Instead, high stakes decisions are made on multiple sources of information. The sampling of behaviour by different assessors in different contexts builds a stable generalisation about a learner's ability, rather than combining all assessments into a single set of examinations.

Programmatic Assessment has a potential role in selection design. This was raised in the 2010 Ottawa statement (Prideaux et al., 2011) and has been referred to by other authors since, usually in their discussion regarding gaps in the literature or areas for future research (Patterson, Lievens, et al., 2013; Roberts et al., 2017). One paper has explored concepts of Programmatic Assessment through mapping domains to selection tools used in medical school selection (Wilkinson & Wilkinson, 2016). Beyond this however, there are no published examples of Programmatic Assessment principles being used as a model to design a selection process.

The Programmatic Assessment framework aims to optimise three functions of assessment programs; learning, curriculum quality assurance and decision-making (van der Vleuten et al., 2014). It is these last two functions that are relevant to selection. The learning function pertains to the role of feedback used to improve learning and performance, described as a shift from assessment *of* learning, to assessment *for* learning (Schuwirth & Van der Vleuten, 2011). While there may be some feedback given to applicants, assessment for learning is not a primary function of selection. Even so, the curriculum design and mapping in Programmatic Assessment has parallels to the blueprinting of attributes and domains to selection tools meaning this is worthy of further consideration, as are the suggested processes for high stakes decision-making.

The alignment of curriculum objectives to assessment tools is a key principle of Programmatic Assessment which is already used in some selection processes. In assessment, first curriculum outcomes and competencies are identified, then assessment tools appropriate for assessing these competencies are chosen. Each assessment provides a single data point about the learner which when combined form a holistic picture. There are several examples in the selection literature (and discussed in the literature review) where desirable attributes or domains are identified and tools are mapped to assess these (Gale et al., 2010; Randall et al., 2006; Roberts et al., 2014). This practice, based on competency based medical education frameworks (Frank et al., 2010), can help ensure a rounded evaluation of applicants as well as reduce redundancy of testing.

Mapping tools to attributes can help coordinators make informed decisions about the design of the selection process. As an example, in the case study the 2015 cohort applicants were asked to reflect on times when they were performing poorly and strategies they use when this occurs, in two separate tasks; by written response in the personal statement and verbally during the interview (Appendix C). Both questions sought similar information and had comparable marking criteria raising the issues of redundancy and efficiency. Through a formal process of mapping measurement tools to domains, coordinators would be able to see that these questions are assessing a similar domain, perhaps ‘self-awareness’ or ‘personal insight’. They may decide this is appropriate as the information is being obtained in different forms (written and oral) and having two forms of the question may act a checking process for consistency of response. Alternatively, coordinators may decide that another domain had not been adequately explored through the process and use the interview to explore that area and avoid repetition. Having an overall master plan can clearly define what data points are collected and allow selection decisions to be made based on performance in different domains.

Making selection decisions based on domains rather than the tools used has been advocated in selection for medical school (Wilkinson and Wilkinson 2016). This concept was explored in a cohort of 507 medical graduates in New Zealand. The selection measures consisted of three sections of the Undergraduate Medical Admissions Test (UMAT) and seven first year university courses. They used correlations between these measures and the score in the final exam in fifth year to group the measures used into specific domains. This was a selection process based around tools. Identification of domains was a retrospective activity that occurred once the tools had been administered and their predictive validity for fifth year exam scores calculated. Analysis found some domains were examined by a single tool, while other domains were assessed across multiple tools. This revealed some potential redundancy with six of the university courses were grouped together under the domain, ‘biomedical science’. They found some selection measures did not correlate with selection exam performance (UMAT Section 3) and questioned their contribution to the overall process. They concluded that the use of redundant measures and those with little contribution to outcome could be avoided using the principles of Programmatic Assessment. This is the only published study to have applied Programmatic Assessment principles to selection, albeit they were applied post hoc. The prospective use of the Programmatic Assessment framework for selection process design is yet to be reported.

Decision-making processes advocated in the Programmatic Assessment model have the potential to address some of the challenges faced in selection. As in selection, assessment panels may have both quantitative and qualitative material to consider when making decisions. The limitations of converting this information into numerical values has already been discussed. For high stakes decisions in assessment, the use of a panel to review all the data points available about a learner is suggested (van der Vleuten et al., 2014). This panel would consider all information available about a learner, performance in the same domain assessed at different time points by different assessors, and then document the reasoning behind the final decision. Rather than the reductionist approach of converting all information to a score, the value of information obtained in written responses, MCQs and supervisor feedback is able to be considered. Could this approach work in selection?

Shared decision-making by a panel who have access to all information on the applicants has several advantages over the currently used compensatory models. Such algorithms for decision-making in selection often allow compensation for low scores on one tool by high scores on another (Randall et al., 2006; Shulruf et al., 2018). Performance in selection tools and hence suitability for selection is averaged out across a range of measures. The suggested panel review approach is a means to address this. In the SWSN case study, an applicant may perform poorly and have low scores for the administration score, the written reflection on clinical handover and the communication exercise. This pattern of poor performance in measures related to communication and professionalism may be hidden from selectors by high scores in clinical knowledge stations. A panel will be able to view all information and make an assessment its relevance to selection outcome. That is not to say all scoring should be abandoned. The panel may set *threshold scores* for individual domains felt to be essential (e.g. all applicants must be able to safely prescribe). Alternatively, they may adopt a *minimum evidence* approach, acknowledging that all domains are considered to have value and must be passed to a minimum level. Failures in one domain cannot be made up for with performance in another. Whatever the approach taken, the panel can tailor this to the context in which the selection process operates.

The scale and timing of selection processes will impact on whether the Programmatic Approach to decision-making can be used. When learners are considered by such a panel, ‘most learners will require very little time; very few will need considerable deliberation’ (van der Vleuten et al., 2014). This is indeed the case in assessment where it is possible for all learners to pass and the reality is most *will* pass. In selection, most applicants will *fail.* Every decision to not select an applicant is significant and requires considerable deliberation. A process like the SWSN case study with a cohort of 32 applicants for 15 positions is likely to be able to make the time to discuss individual applicants. For larger cohorts like those seen in medical school selection there are concerns that this would not be feasible (Prideaux et al., 2011).

Another potential barrier to making decisions via a Programmatic Assessment review panel is the social validity of such a process. The integrity of assessment decisions is thought to be enhanced by the panel review process as it 'will usually lead to robust decisions that have credibility and can be trusted' (van der Vleuten et al., 2014). The implication is that because a group of faculty have met and discussed the case, the outcome has some veracity that is superior to adding the sum of assessment scores. In the case study there was some perceived assurance offered by basing decisions on scores. The numerical scoring system was viewed as defensible. When unsuccessful applicants sought feedback, communicating that they had a lower score than successful applicants, was a less complicated discussion than having to address the specific reasons as to why that was the case. One of the instigating factors for the Brennan Review in medical specialty selection was concern about the transparency of selection processes (Brennan, 1998). Selection coordinators in this case study shared concerns about appearing transparent:

> *There's no secret secondary process where, "Oh my God, how did we select that person?" [laughs] … This whole effort is to stop that. (Interviewee 2, p30-31)*

Ironically, converting qualitative data into numerical scores could be considered the antithesis of transparent. To trust in a numerical scoring system over the value of subjective human judgement is the contradictory view to that promoted in Programmatic Assessment.

The other reason to value the use of numerical scoring is a practical one, the imperative to rank applicants for selection. Unlike medical education assessment where every learner can pass, suitable competent applicants will not be selected. In this context of sorting competent applicants, a ranking system of some sort is needed to offer positions to successful applicants, and then subsequently to make further offers when there are withdrawals. A review panel may well be able to deliberate about several competent applicants and place them in rank order however a numerical scoring system provides an efficient and pragmatic process for this to occur.

Programmatic Assessment provides a guiding framework for selection design that addresses some of the challenges reported and demonstrated in this thesis in delivering a fair and effective selection process. The principles of Programmatic Assessment could potentially enhance the alignment of tools to selection domains and also facilitate decision-making based on richer information. The challenge in using a programmatic approach is how this might provide a means to defensibly distinguish between the selected and those not selected - given all applicants may be qualified for selection. Even if all aspects of a programmatic approach are optimised; i.e. domains are carefully described and mapped to selection tools, multiple data points are available on applicants and this information (both quantitative and qualitative), is available to a decision-making panel, who use algorithms that accommodate threshold scores and minimal evidence of achievement in all domains – a process to discriminate amongst several suitable applicants is needed. Whether a Programmatic Assessment approach to selection can achieve this is yet to be tested. Future research should consider using programmatic assessment to design a selection process and investigate if decision-making processes are feasible to discriminate and rank applicants.

7. CONCLUSION

This thesis set out to explore the question:

What is an effective process for selecting applicants into vocational medical training programs?

Determining whether a process is effective proved to be a challenge. The most important measure of the effectiveness of selection processes is whether the successful applicants go on to become successful trainees and specialists. We do not have meaningful methods to define and measure success in training. Quality research on predictive validity in vocational medical selection is not currently available and due to the difficulties outlined earlier (p32-33), may never be. In this vacuum, attention has been given to optimising the reliability of individual measurement tools and also reporting on the political validity of selection processes. The tools described in the literature are *methods* of collecting data on applicants, not *measurements* themselves, and their reliability, construct validity, utility, and thus the effectiveness, is influenced by many factors. Conceptual frameworks for designing, implementing and evaluating selection processes are an emerging area of research. Presently there is no clear accepted standard for an effective process in vocational medical selection.

Through analysing a local case study of the SWSN, the challenges of vocational selection were able to be explored in detail. Selection coordinators of this process gave great emphasis to the tools used and their content, and sought to change the scoring systems and format of the tools to enhance fairness and facilitate logistics. Valuable information on each applicant was collected and converted to a score used to make decisions about who should be selected. The decision-making algorithm used meant the weighting given to each tool became the most influential factor in determining who was selected. There was tension in the decision-making process with concern whether variances in scores reflected meaningful differences in applicants.

Both the literature review and case study revealed the need for a framework with which to design selection. The case study provided specific examples through which to discuss the model of Programmatic Assessment. The central key is that a *program of selection* would allow the whole picture of an applicant's competence to be obtained by a careful choice of selection methods, and a structured plan about how data from tools are combined to make decisions. There is the potential for the principles of Programmatic Assessment to assist with the mapping of domains to tools used in selection and also in facilitating decision-making processes making full use of all information gathered about an applicant. However meaningful discrimination between several highly competent applicants is still difficult – a challenge that remains unresolved. Future research should explore using programmatic assessment principles to design a selection process, with particular attention to combining information to make final selection decisions.

Ultimately, what makes a selection process effective is multi-factorial and difficult to measure. Meaningful predictive validity studies are the holy grail in selection and yet may never be able to be undertaken, so selection processes in vocational medical training need to be viewed from a deontological perspective. That is, a process is needed to select trainees that is fair, has social and political validity to key stakeholders and is psychometrically strong. It is important to have a quality process in place that aspires to have an effective outcome – even if this outcome is hard to measure. Achieving all these goals is a challenge and requires trade-offs between the psychometric rigor of the tools, the views of stakeholders and boundaries of available resources. The principles of Programmatic Assessment may serve as a useful framework to plan a selection process and consider the compromises that need to be made. Prospective use of this framework will reveal whether it is a feasible model for designing an effective selection process.

REFERENCES

Adam, J., Bore, M., Childs, R., Dunn, J., McKendree, J., Munro, D., & Powis, D. (2017). Response to: 'How effective are selection methods in medical education? A systematic review'. *Med Educ*. doi:10.1111/medu.13243

Adams, D., Sice, P., Anderson, I., Gale, T., Lam, H., Langton, J., . . . Carr, A. (2009). Validation of simulation for recruitment to training posts in anaesthesia. *Anaesthesia, 64*(7), 805-806. doi:10.1111/j.1365-2044.2009.05966_18.x

Adusumilli, S., Cohan, R. H., Marshall, K. W., Fitzgerald, J. T., Oh, M. S., Gross, B. H., & Ellis, J. H. (2000). How well does applicant rank order predict subsequent performance during radiology residency? *Acad Radiol, 7*(8), 635-640.

Anderson, N. (2011). Perceived Job Discrimination: Toward a model of applicant propensity to case initiation in selection. *International Journal of Selection and Assessment, 19*(3), 229-244.

Baker, J. D., 3rd, Wallace, C. T., Cooke, J. E., Alpert, C. C., & Ackerly, J. A. (1987). Selection of anesthesiology residents.[Erratum appears in South Med J 1988 May;81(5):683]. *Southern Medical Journal, 80*(8), 1031-1035.

Bandiera, G., & Regehr, G. (2003). Evaluation of a structured application assessment instrument for assessing applications to Canadian postgraduate training programs in emergency medicine. *Academic Emergency Medicine, 10*(6), 594-598.

Bandiera, G., & Regehr, G. (2004). Reliability of a structured interview scoring instrument for a Canadian postgraduate emergency medicine training program. *Acad Emerg Med, 11*(1), 27-32.

Barrett, A., Galvin, R., Steinert, Y., Scherpbier, A., O'Shaughnessy, A., Horgan, M., & Horsley, T. (2015). A BEME (Best Evidence in Medical Education) systematic review of the use of workplace-based assessment in identifying and remediating poor performance among postgraduate medical trainees. *Syst Rev, 4*, 65. doi:10.1186/s13643-015-0056-9

Bell, J. G., Kanellitsas, I., & Shaffer, L. (2002). Selection of obstetrics and gynecology residents on the basis of medical school performance. *Am J Obstet Gynecol, 186*(5), 1091-1094.

Beskind, D. L., Hiller, K. M., Stolz, U., Bradshaw, H., Berkman, M., Stoneking, L. R., . . . Grall, K. J. (2014). Does the experience of the writer affect the evaluative components on the standardized letter of recommendation in emergency medicine? *J Emerg Med, 46*(4), 544-550. doi:10.1016/j.jemermed.2013.08.025

Brennan, P. (1998). *Trainee Selection in Australian Medical Colleges*. Retrieved from Canberra ACT: https://www.surgeons.org/media/20984208/rpt_brennan_report_1998.pdf

Burgess, A., Roberts, C., Clark, T., & Mossman, K. (2014). The social validity of a national assessment centre for selection into general practice training. *BMC Med Educ, 14*(1), 261. doi:10.1186/s12909-014-0261-6

Burgess, A., Roberts, C., Sureshkumar, P., & Mossman, K. (2018). Multiple mini interview (MMI) for general practice training selection in Australia: interviewers' motivation. *BMC Med Educ, 18*(1), 21. doi:10.1186/s12909-018-1128-z

Chu, S., Kaider, A., & Johnson, L. (2017). Selection into Emergency Medicine specialist training: A commentary on the science of selection. *Emerg Med Australas, 29*(4), 461-463. doi:10.1111/1742-6723.12827

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86*(3), 410-417. doi:http://psycnet.apa.org/doi/10.1037/0021-9010.86.3.410

Colliver, J. A., Conlee, M. J., & Verhulst, S. J. (2012). From test validity to construct validity ... and back? *Med Educ, 46*(4), 366-371. doi:10.1111/j.1365-2923.2011.04194.x

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med, 119*(2), 166 e167-116. doi:10.1016/j.amjmed.2005.10.036

Dawkins, K., Ekstrom, R. D., Maltbie, A., & Golden, R. N. (2005). The relationship between psychiatry residency applicant evaluations and subsequent residency performance. *Acad Psychiatry, 29*(1), 69-75. doi:10.1176/appi.ap.29.1.69

de Vet, H. C., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006). When to use agreement versus reliability measures. *J Clin Epidemiol, 59*(10), 1033-1039. doi:10.1016/j.jclinepi.2005.10.015

Dore, K. L., Kreuger, S., Ladhani, M., Rolfson, D., Kurtz, D., Kulasegaram, K., . . . Reiter, H. I. (2010). The reliability and acceptability of the Multiple Mini-Interview as a selection instrument for postgraduate admissions. *Acad Med, 85*(10 Suppl), S60-63. doi:10.1097/ACM.0b013e3181ed442b

Downing, S. M. (2004). Reliability: on the reproducibility of assessment data. *Med Educ, 38*(9), 1006-1012. doi:10.1111/j.1365-2929.2004.01932.x

Ende, J. (1983). Feedback in clinical medical education. *JAMA, 250*(6), 777-781.

Eva, K. W., Macala, C., & Fleming, B. (2018). Twelve tips for constructing a multiple mini-interview. *Med Teach*, 1-7. doi:10.1080/0142159X.2018.1429586

Eva, K. W., Reiter, H. I., Trinh, K., Wasi, P., Rosenfeld, J., & Norman, G. R. (2009). Predictive validity of the multiple mini-interview for selecting medical trainees. *Med Educ, 43*(8), 767-775. doi:10.1111/j.1365-2923.2009.03407.x

Ferguson, E., James, D., O'Hehir, F., Sanders, A., & McManus, I. C. (2003). Pilot study of the roles of personality, references, and personal statements in relation to performance over the five years of a medical degree. *BMJ, 326*(7386), 429-432. doi:10.1136/bmj.326.7386.429

Frank, J. R., Mungroo, R., Ahmad, Y., Wang, M., De Rossi, S., & Horsley, T. (2010). Toward a definition of competency-based education in medicine: a systematic review of published definitions. *Med Teach, 32*(8), 631-637. doi:10.3109/0142159X.2010.500898

Frumovitz, M., Kriseman, M. L., Sun, C. C., Blumenthal-Barby, J., Sood, A. K., Bodurka, D. C., & Soliman, P. T. (2012). Unverifiable accomplishments and publications on applications for gynecologic oncology fellowships. *Obstetrics and Gynecology, 119*(3), 504-508. doi:10.1097/AOG.0b013e31824206e9

Gale, T. C. E., Roberts, M. J., Sice, P. J., Langton, J. A., Patterson, F. C., Carr, A. S., . . . Davies, P. R. F. (2010). Predictive validity of a selection centre testing non-technical skills for recruitment to training in anaesthesia. *British Journal of Anaesthesia, 105*(5), 603-609. doi:10.1093/bja/aeq228

Goodwin, M., Large, D., Kerrin, M., Honsberger, J., Carr, A., & Wilkinson, D. (2014). Developing national selection processes for entry into postgraduate specialty training: the case of trauma and orthopedics in the United Kingdom. *Curr Rev Musculoskelet Med, 7*(2), 145-150. doi:10.1007/s12178-014-9206-2

Goodyear, H. M., Jyothish, D., Diwakar, V., & Wall, D. (2007). Reliability of a regional junior doctor recruitment process. *Med Teach, 29*(5), 504-506. doi:10.1080/01421590701526823

Grantcharov, T. P., & Reznick, R. K. (2009). Training tomorrow's surgeons: what are we looking for and how can we achieve it? *ANZ J Surg, 79*(3), 104-107. doi:10.1111/j.1445-2197.2008.04823.x

Hodges, B. (2013). Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach, 35*(7), 564-568. doi:10.3109/0142159X.2013.789134

Hofmeister, M., Lockyer, J., & Crutcher, R. (2009). The multiple mini-interview for selection of international medical graduates into family medicine residency education. *Medical Education, 43*(6), 573-579. doi:10.1111/j.1365-2923.2009.03380.x

Kelly, M. E., Patterson, F., O'Flynn, S., Mulligan, J., & Murphy, A. W. (2018). A systematic review of stakeholder views of selection methods for medical schools admission. *BMC Med Educ, 18*(1), 139. doi:10.1186/s12909-018-1235-x

Knorr, M., & Hissbach, J. (2014). Multiple mini-interviews: same concept, different approaches. *Med Educ, 48*(12), 1157-1175. doi:10.1111/medu.12535

Koczwara, A., Patterson, F., Zibarras, L., Kerrin, M., Irish, B., & Wilkinson, M. (2012). Evaluating cognitive ability, knowledge tests and situational judgement tests for postgraduate selection. *Med Educ, 46*(4), 399-408. doi:10.1111/j.1365-2923.2011.04195.x

Kuper, A., Reeves, S., Albert, M., & Hodges, B. D. (2007). Assessment: do we need to broaden our methodological horizons? *Med Educ, 41*(12), 1121-1123. doi:10.1111/j.1365-2923.2007.02945.x

Laschinger, H. K. (1992). Intraclass correlations as estimates of interrater reliability in nursing research. *Western Journal of Nursing Research, 14*(2), 246-251.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: illustrating the importance of matching predictor and criterion construct domains. *J Appl Psychol, 90*(3), 442-452. doi:10.1037/0021-9010.90.3.442

Makdisi, G., Takeuchi, T., Rodriguez, J., Rucinski, J., & Wise, L. (2011). How we select our residents--a survey of selection criteria in general surgery residents. *J Surg Educ, 68*(1), 67-72. doi:10.1016/j.jsurg.2010.10.003

Medical Deans Australia and New Zealand. (2015). *Workforce Data Report 2015*. Retrieved from

Medical Training Review Panel. (2015). *Medical Training Review Panel Eighteenth Report*. Retrieved from

Merriam, S. B. (1998). Qualitative research and case study applications in education. In S. B. Merriam (Ed.), (2nd ed. ed.). San Francisco :: Jossey-Bass Publishers.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement*. New York: American Council on Education and Macmillian.

Mitchison, H. (2009). Assessment centres for core medical training: how do the assessors feel this compares with the traditional interview? *Clin Med, 9*(2), 147-150.

Newble, D. I., Baxter, A., & Elmslie, R. G. (1979). A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. *Med Educ, 13*(4), 263-268.

Olawaiye, A., Yeh, J., & Withiam-Leitch, M. (2006). Resident selection process and prediction of clinical performance in an obstetrics and gynecology program. *Teach Learn Med, 18*(4), 310-315. doi:10.1207/s15328015tlm1804_6

Oldfield, Z. B., Spencer W. Smith, Julian Anthony, Adrian Watt, Anthony. (2013). Correlation of selection scores with subsequent assessment scores during surgical training. *ANZ Journal of Surgery, 83*(6), 412-416.

Pashayan, N., Gray, S., Duff, C., Parkes, J., Williams, D., Patterson, F., . . . Mason, B. W. (2015). Evaluation of recruitment and selection for specialty training in public health: interim results of a prospective cohort study to measure the predictive validity of the selection process. *J Public Health (Oxf)*. doi:10.1093/pubmed/fdv102

Patterson, F., Ashworth, V., Zibarras, L., Coan, P., Kerrin, M., & O'Neill, P. (2012). Evaluations of situational judgement tests to assess non-academic attributes in selection. *Med Educ, 46*(9), 850-868. doi:10.1111/j.1365-2923.2012.04336.x

Patterson, F., Cleland, J., & Cousans, F. (2017). Selection methods in healthcare professions: where are we now and where next? *Adv Health Sci Educ Theory Pract, 22*(2), 229-242. doi:10.1007/s10459-017-9752-7

Patterson, F., & Ferguson, E. (2012). Testing non-cognitive attributes in selection centres: how to avoid being reliably wrong. *Med Educ, 46*(3), 240-242. doi:10.1111/j.1365-2923.2011.04193.x

Patterson, F., Ferguson, E., & Knight, A. (2013). Selection into medical education and training. In T. Swanwick (Ed.), *Understanding Medical Education: Evidence, Theory and Practice* (pp. 403-420): Wiley Blackwell.

Patterson, F., Ferguson, E., & Thomas, S. (2008). Using job analysis to identify core and specific competencies: implications for selection and recruitment. *Med Educ, 42*(12), 1195-1204. doi:10.1111/j.1365-2923.2008.03174.x

Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., & Cleland, J. (2016). How effective are selection methods in medical education? A systematic review. *Med Educ, 50*(1), 36-60. doi:10.1111/medu.12817

Patterson, F., Lievens, F., Kerrin, M., Munro, N., & Irish, B. (2013). The predictive validity of selection for entry into postgraduate training in general practice: evidence from three longitudinal studies. *Br J Gen Pract, 63*(616), e734-741. doi:10.3399/bjgp13X674413

Patterson, F., Lievens, F., Kerrin, M., Zibarras, L., & Carette, B. (2012). Designing Selection Systems for Medicine: The importance of balancing predictive and political validity in high-stakes selection contexts. *International Journal of Selection and Assessment, 20*(4), 486-496.

Patterson, F., Rowett, E., Hale, R., Grant, M., Roberts, C., Cousans, F., & Martin, S. (2016). The predictive validity of a situational judgement test and multiple-mini interview for entry into postgraduate training in Australia. *BMC Med Educ, 16*(1), 87. doi:10.1186/s12909-016-0606-4

Patterson, F., Tavabie, A., Denney, M., Kerrin, M., Ashworth, V., Koczwara, A., & MacLeod, S. (2013). A new competency model for general practice: implications for selection, training, and careers. *British Journal of General Practice, 63*(610), e331-338.

Patterson, F., Zibarras, L., & Ashworth, V. (2016). Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100. *Med Teach, 38*(1), 3-17. doi:10.3109/0142159X.2015.1072619

Prager, J. D., Myer, C. M., 3rd, & Pensak, M. L. (2010). Improving the letter of recommendation. *Otolaryngol Head Neck Surg, 143*(3), 327-330. doi:10.1016/j.otohns.2010.03.017

Prideaux, D., Roberts, C., Eva, K., Centeno, A., McCrorie, P., McManus, C., . . . Wilkinson, D. (2011). Assessment for selection for the health care professions and specialty training: consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach, 33*(3), 215-223. doi:10.3109/0142159X.2011.551560

Quintero, A. J., Segal, L. S., King, T. S., & Black, K. P. (2009). The personal interview: assessing the potential for personality similarity to bias the selection of orthopaedic residents. *Acad Med, 84*(10), 1364-1372. doi:10.1097/ACM.0b013e3181b6a9af

Randall, R., Davies, H., Patterson, F., & Farrell, K. (2006). Selecting doctors for postgraduate training in paediatrics using a competency based assessment centre. *Archives of Disease in Childhood, 91*(5), 444-448.

Razack, S., Faremo, S., Drolet, F., Snell, L., Wiseman, J., & Pickering, J. (2009). Multiple mini-interviews versus traditional interviews: stakeholder acceptability comparison. *Med Educ, 43*(10), 993-1000. doi:10.1111/j.1365-2923.2009.03447.x

Roberts, C., Clark, T., Burgess, A., Frommer, M., Grant, M., & Mossman, K. (2014). The validity of a behavioural multiple-mini-interview within an assessment centre for selection into specialty training. *BMC Med Educ, 14*, 169.

Roberts, C., Khanna, P., Rigby, L., Bartle, E., Llewellyn, A., Gustavs, J., . . . Lynam, J. (2017). Utility of selection methods for specialist medical training: A BEME (best evidence medical education) systematic review: BEME guide no. 45. *Med Teach*, 1-17. doi:10.1080/0142159X.2017.1367375

Roberts, C., & Togno, J. M. (2011). Selection into specialist training programs: an approach from general practice. *Medical Journal of Australia, 194*(2), 93-95.

Royal Australasian College of Physicians. (2018). Selection Into Training. Retrieved from https://www.racp.edu.au/innovation/education-renewal/selection-into-training

Schmidt, H. G., Norman, G. R., & Boshuizen, H. P. (1990). A cognitive perspective on medical expertise: theory and implication. *Acad Med, 65*(10), 611-621.

Schoenherr, J. R., & Hamstra, S. J. (2016). Psychometrics and its discontents: an historical perspective on the discourse of the measurement tradition. *Adv Health Sci Educ Theory Pract, 21*(3), 719-729. doi:10.1007/s10459-015-9623-z

Schuler HI, S. H., Farr JL, Smith M. (1993). *Social validity of selection situations: A concept and some empirical results*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Schuwirth, L. W., & van der Vleuten, C. P. (2006). A plea for new psychometric models in educational assessment. *Med Educ, 40*(4), 296-300. doi:10.1111/j.1365-2929.2006.02405.x

Schuwirth, L. W., & Van der Vleuten, C. P. (2011). Programmatic assessment: From assessment of learning to assessment for learning. *Med Teach, 33*(6), 478-485. doi:10.3109/0142159X.2011.565828

Selber, J. C., Tong, W., Koshy, J., Ibrahim, A., Liu, J., & Butler, C. (2014). Correlation between trainee candidate selection criteria and subsequent performance. *J Am Coll Surg, 219*(5), 951-957. doi:10.1016/j.jamcollsurg.2014.07.942

Shulruf, B., Bagg, W., Begun, M., Hay, M., Lichtwark, I., Turnock, A., . . . Poole, P. J. (2018). The efficacy of medical student selection tools in Australia and New Zealand. *Med J Aust, 208*(5), 214-218.

Soares, W. E., 3rd, Sohoni, A., Hern, H. G., Wills, C. P., Alter, H. J., & Simon, B. C. (2015). Comparison of the multiple mini-interview with the traditional interview for U.S. emergency medicine residency applicants: a single-institution experience. *Acad Med, 90*(1), 76-81. doi:10.1097/ACM.0000000000000524

Stedman, J. M., Hatch, J. P., & Schoenfeld, L. S. (2009). Letters of recommendation for the predoctoral internship in medical schools and other settings: do they enhance decision making in the selection process? *J Clin Psychol Med Settings, 16*(4), 339-345. doi:10.1007/s10880-009-9170-y

Swanson, W. S., Harris, M. C., Master, C., Gallagher, P. R., Mauro, A. E., & Ludwig, S. (2005). The impact of the interview in pediatric residency selection. *Ambul Pediatr, 5*(4), 216-220. doi:10.1367/A04-149R1.1

Tavakol, M., & Sandars, J. (2014). Quantitative and qualitative methods in medical education research: AMEE Guide No 90: Part I. *Med Teach, 36*(9), 746-756. doi:10.3109/0142159X.2014.915298

Thomas, J. (2017). Selection into Emergency Medicine specialist training: The art of selection. *Emerg Med Australas, 29*(4), 459-460. doi:10.1111/1742-6723.12826 van der Vleuten, C. P. (1996). The assessment of professional competence: Developments, research and practical implications. *Adv Health Sci Educ Theory Pract, 1*(1), 41-67. doi:10.1007/BF00596229

van der Vleuten, C. P. (2016). Revisiting 'Assessing professional competence: from methods to programmes'. *Med Educ, 50*(9), 885-888. doi:10.1111/medu.12632

van der Vleuten, C. P., & Schuwirth, L. W. (2005). Assessing professional competence: from methods to programmes. *Med Educ, 39*(3), 309-317. doi:10.1111/j.1365-2929.2005.02094.x

van der Vleuten, C. P., Schuwirth, L. W., Driessen, E. W., Govaerts, M. J., & Heeneman, S. (2014). 12 Tips for programmatic assessment. *Med Teach*, 1-6. doi:10.3109/0142159X.2014.973388

Vermeulen, M. I., Tromp, F., Zuithoff, N. P. A., Pieters, R. H. M., Damoiseaux, R. A. M. J., & Kuyvenhoven, M. M. (2014). A competency based selection procedure for Dutch postgraduate GP training: A pilot study on validity and reliability. *European Journal of General Practice, 20*(4), 307-313. doi:10.3109/13814788.2014.885013

Watson, P. F., & Petrie, A. (2010). Method agreement analysis: a review of correct methodology. *Theriogenology, 73*(9), 1167-1179. doi:10.1016/j.theriogenology.2010.01.003

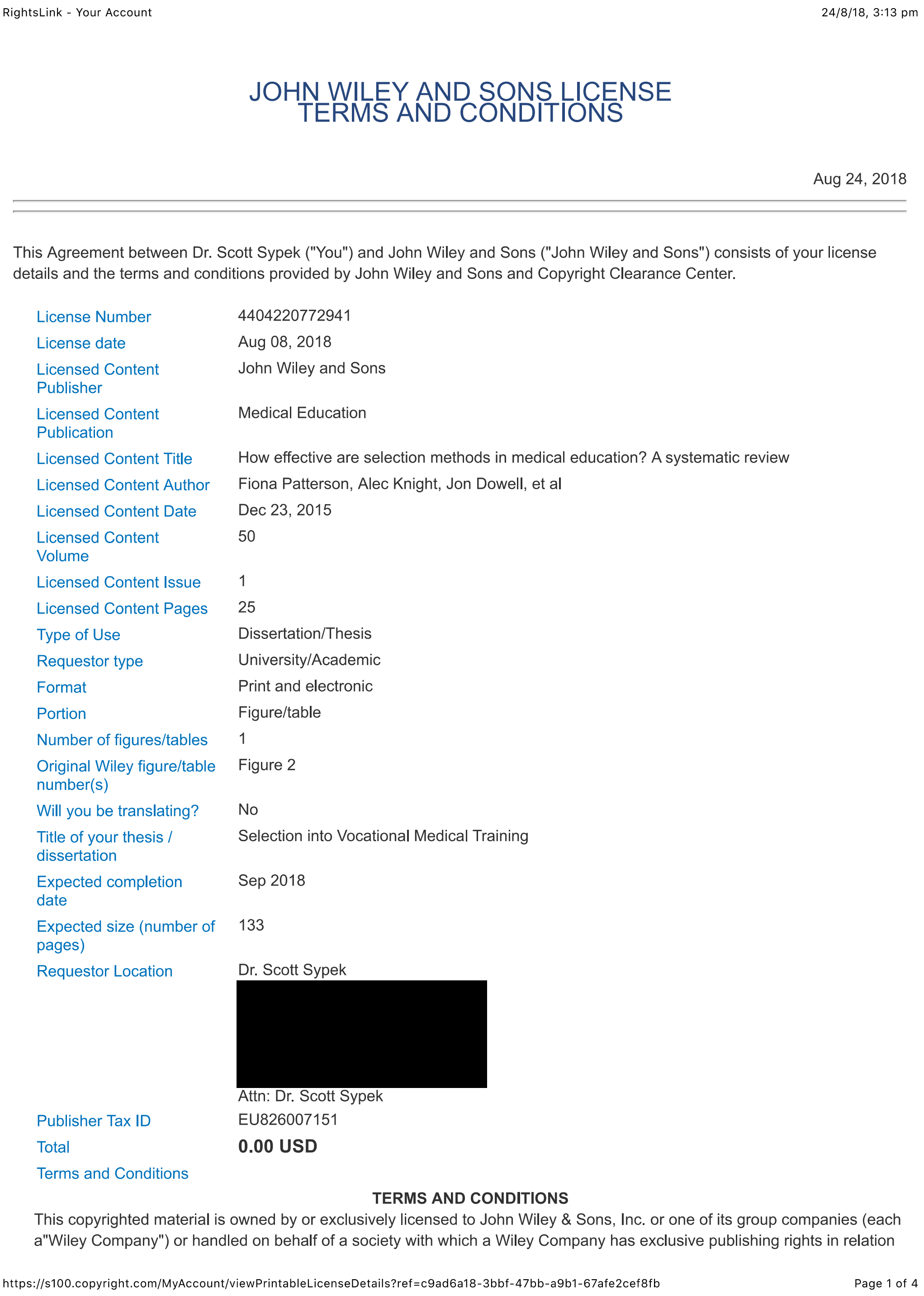
Wilkinson, T. M., & Wilkinson, T. J. (2016). Selection into medical school: from tools to domains. *BMC Med Educ, 16*(1), 258. doi:10.1186/s12909-016-0779-x

Yin, R. (2009). *Case Study Research: Design and Methods* (4th ed. Vol. 5). California: SAGE.

Yoshimura, H., Kitazono, H., Fujitani, S., Machi, J., Saiki, T., Suzuki, Y., & Ponnamperuma, G. (2015). Past-behavioural versus situational questions in a postgraduate admissions multiple mini-interview: a reliability and acceptability comparison. *BMC Med Educ, 15*, 75. doi:10.1186/s12909-015-0361-y

APPENDICES

Appendix A Patterson (2016) Copyright License Agreement



JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

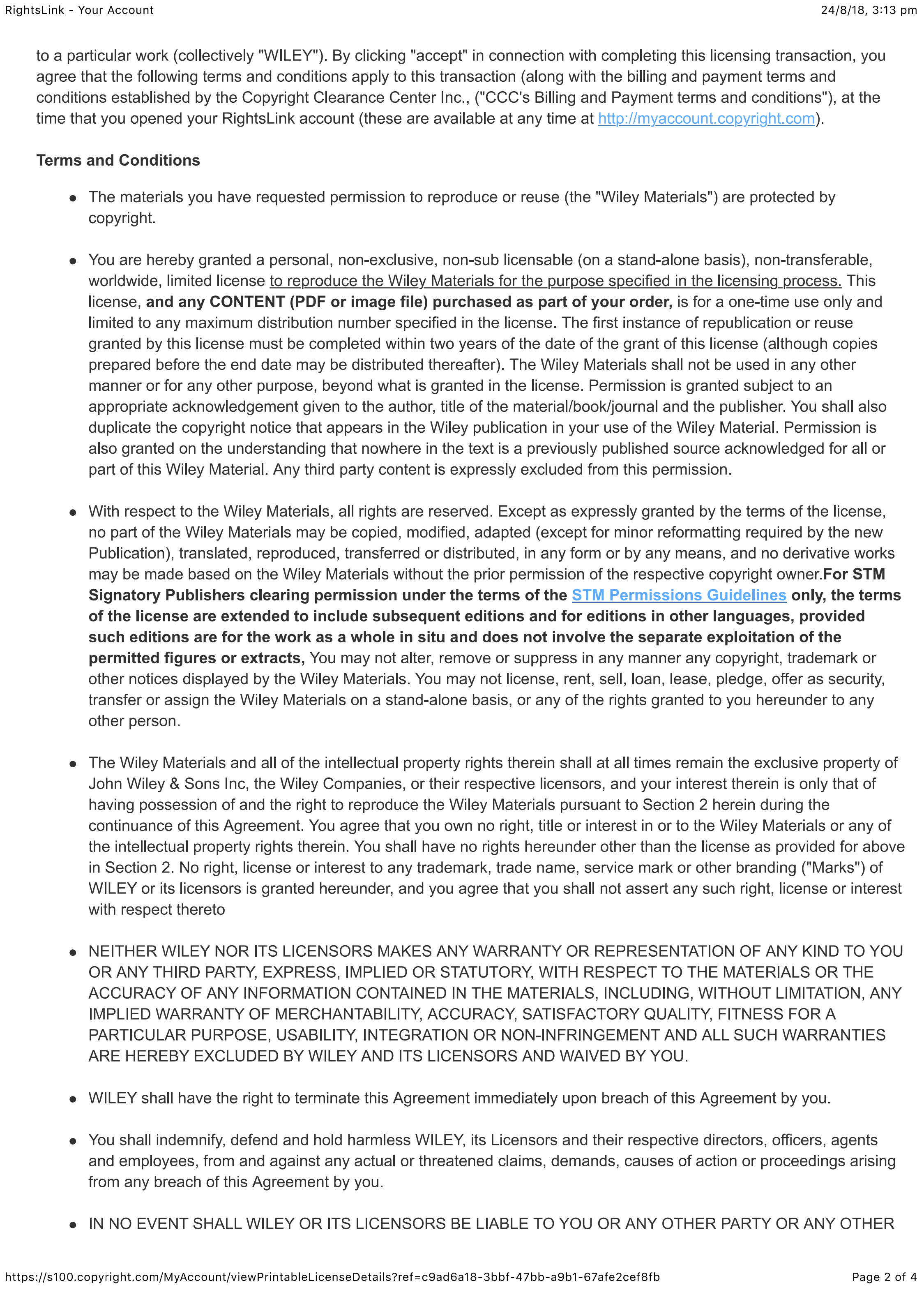
Aug 24, 2018

This Agreement between Dr. Scott Sypek ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	4404220772941
License date	Aug 08, 2018
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Medical Education
Licensed Content Title	How effective are selection methods in medical education? A systematic review
Licensed Content Author	Fiona Patterson, Alec Knight, Jon Dowell, et al
Licensed Content Date	Dec 23, 2015
Licensed Content Volume	50
Licensed Content Issue	1
Licensed Content Pages	25
Type of Use	Dissertation/Thesis
Requestor type	University/Academic
Format	Print and electronic
Portion	Figure/table
Number of figures/tables	1
Original Wiley figure/table number(s)	Figure 2
Will you be translating?	No
Title of your thesis / dissertation	Selection into Vocational Medical Training
Expected completion date	Sep 2018
Expected size (number of pages)	133
Requestor Location	Dr. Scott Sypek [redacted] Attn: Dr. Scott Sypek
Publisher Tax ID	EU826007151
Total	**0.00 USD**
Terms and Conditions	

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation



to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.

- You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, **and any CONTENT (PDF or image file) purchased as part of your order,** is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the grant of this license (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.

- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner.**For STM Signatory Publishers clearing permission under the terms of the STM Permissions Guidelines only, the terms of the license are extended to include subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and does not involve the separate exploitation of the permitted figures or extracts,** You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.

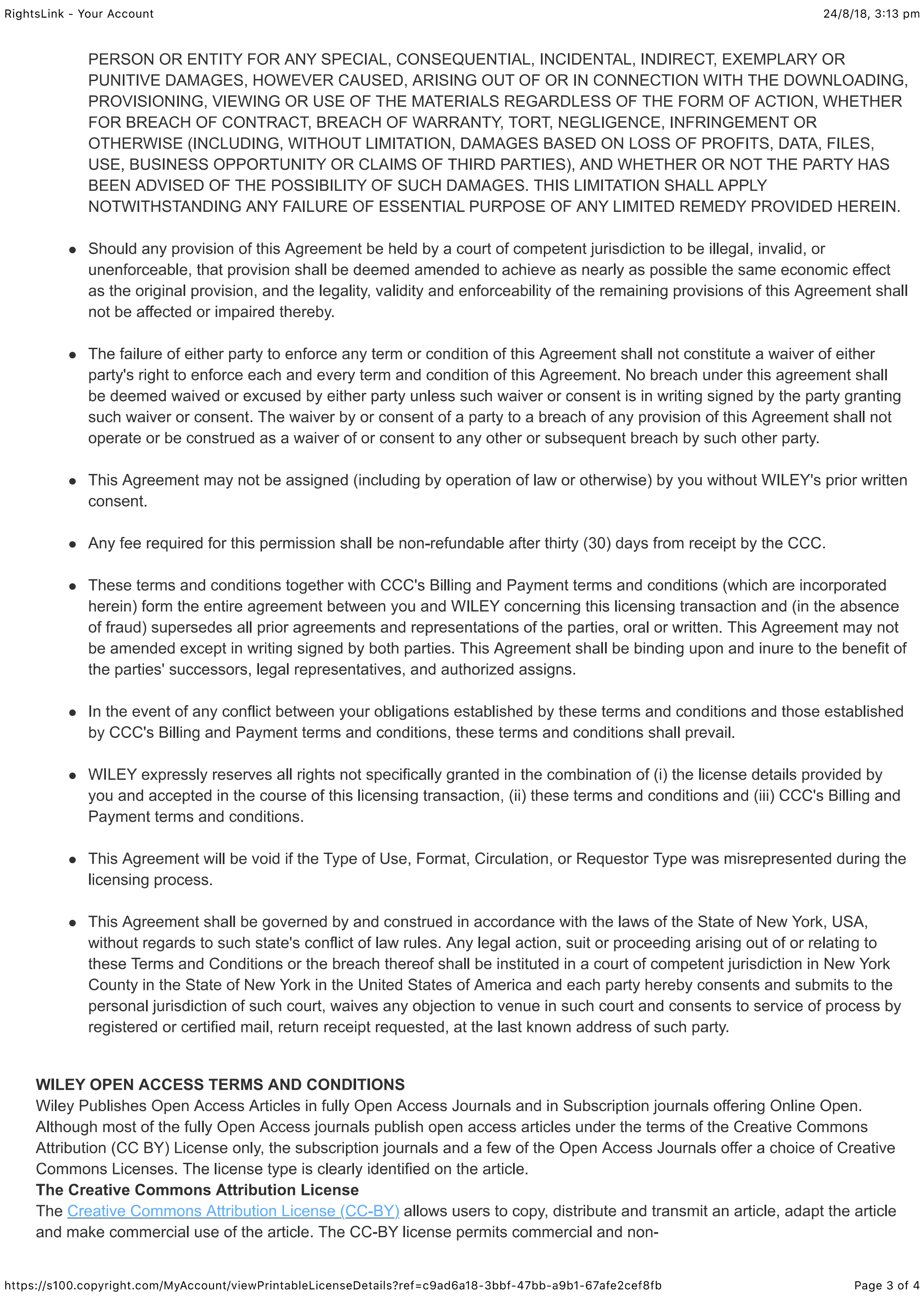
- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto

- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER



PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.

- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

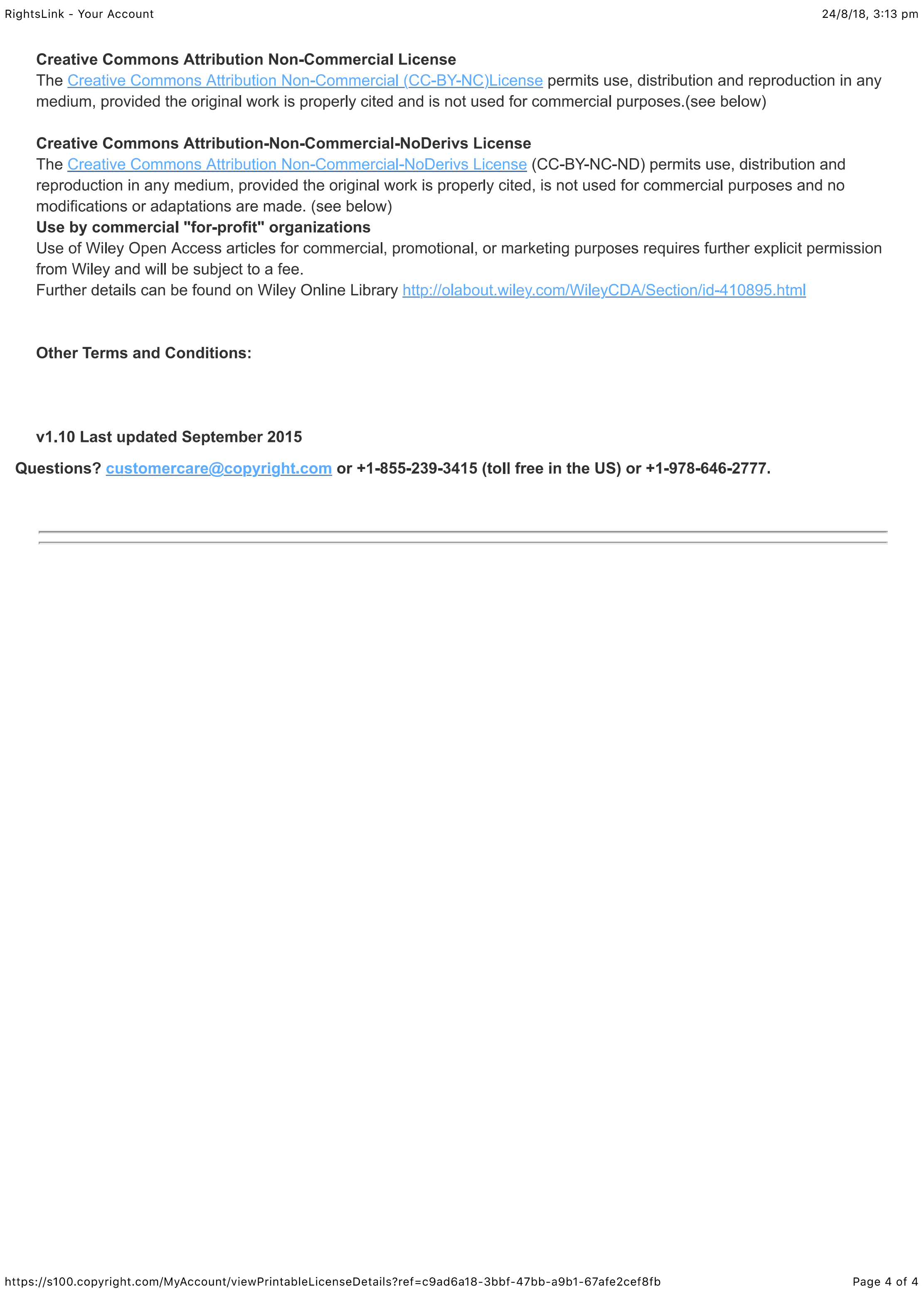
- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS

Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses. The license type is clearly identified on the article.

The Creative Commons Attribution License

The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-



Creative Commons Attribution Non-Commercial License

The Creative Commons Attribution Non-Commercial (CC-BY-NC)License permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.(see below)

Creative Commons Attribution-Non-Commercial-NoDerivs License

The Creative Commons Attribution Non-Commercial-NoDerivs License (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)

Use by commercial "for-profit" organizations

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee.

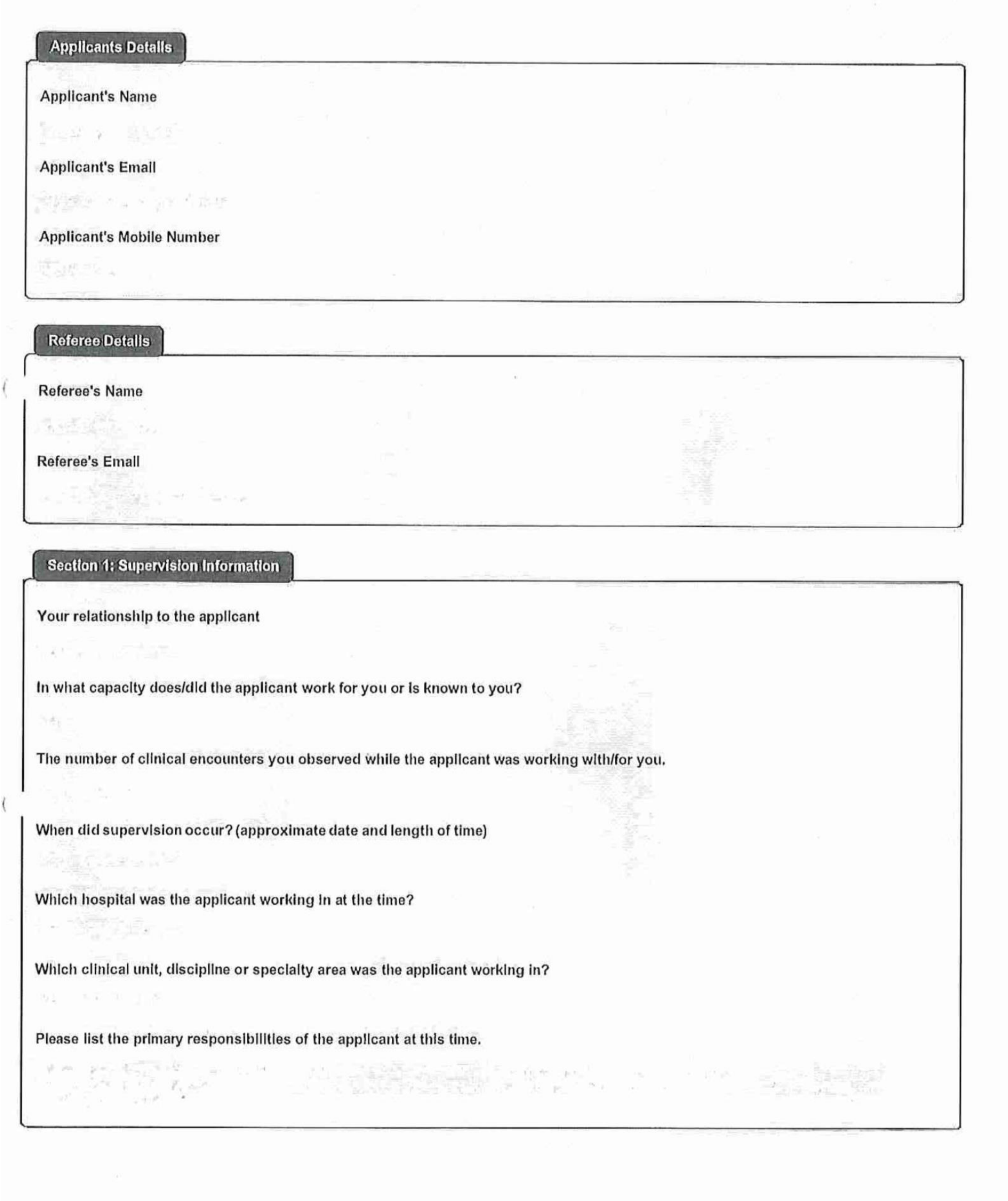
Further details can be found on Wiley Online Library http://olabout.wiley.com/WileyCDA/Section/id-410895.html

Other Terms and Conditions:

v1.10 Last updated September 2015

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Appendix B Example referee report form



Applicants Details

Applicant's Name

Applicant's Email

Applicant's Mobile Number

Referee Details

Referee's Name

Referee's Email

Section 1: Supervision Information

Your relationship to the applicant

In what capacity does/did the applicant work for you or is known to you?

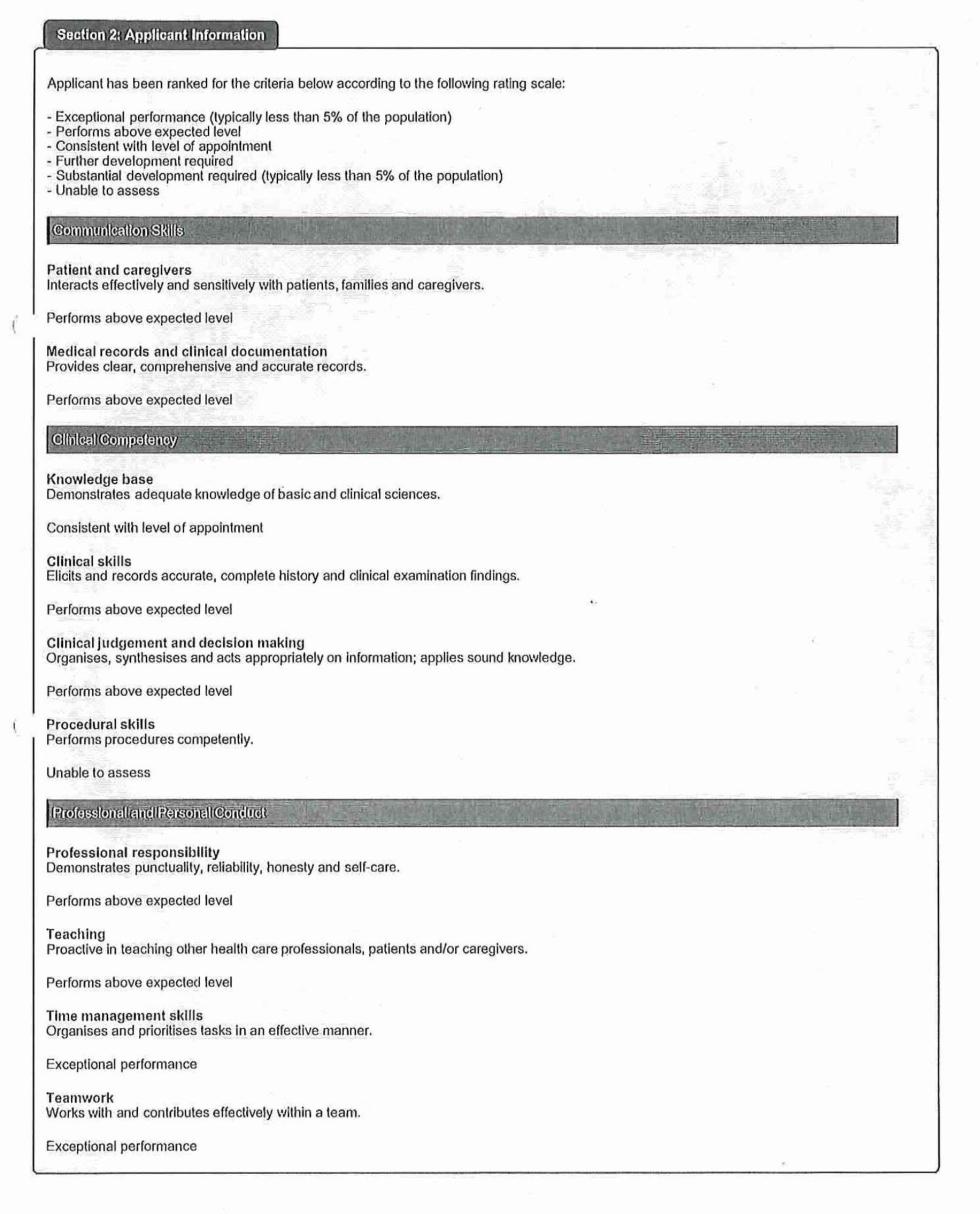
The number of clinical encounters you observed while the applicant was working with/for you.

When did supervision occur? (approximate date and length of time)

Which hospital was the applicant working in at the time?

Which clinical unit, discipline or specialty area was the applicant working in?

Please list the primary responsibilities of the applicant at this time.



Section 2: Applicant Information

Applicant has been ranked for the criteria below according to the following rating scale:

- Exceptional performance (typically less than 5% of the population)
- Performs above expected level
- Consistent with level of appointment
- Further development required
- Substantial development required (typically less than 5% of the population)
- Unable to assess

Communication Skills

Patient and caregivers
Interacts effectively and sensitively with patients, families and caregivers.

Performs above expected level

Medical records and clinical documentation
Provides clear, comprehensive and accurate records.

Performs above expected level

Clinical Competency

Knowledge base
Demonstrates adequate knowledge of basic and clinical sciences.

Consistent with level of appointment

Clinical skills
Elicits and records accurate, complete history and clinical examination findings.

Performs above expected level

Clinical judgement and decision making
Organises, synthesises and acts appropriately on information; applies sound knowledge.

Performs above expected level

Procedural skills
Performs procedures competently.

Unable to assess

Professional and Personal Conduct

Professional responsibility
Demonstrates punctuality, reliability, honesty and self-care.

Performs above expected level

Teaching
Proactive in teaching other health care professionals, patients and/or caregivers.

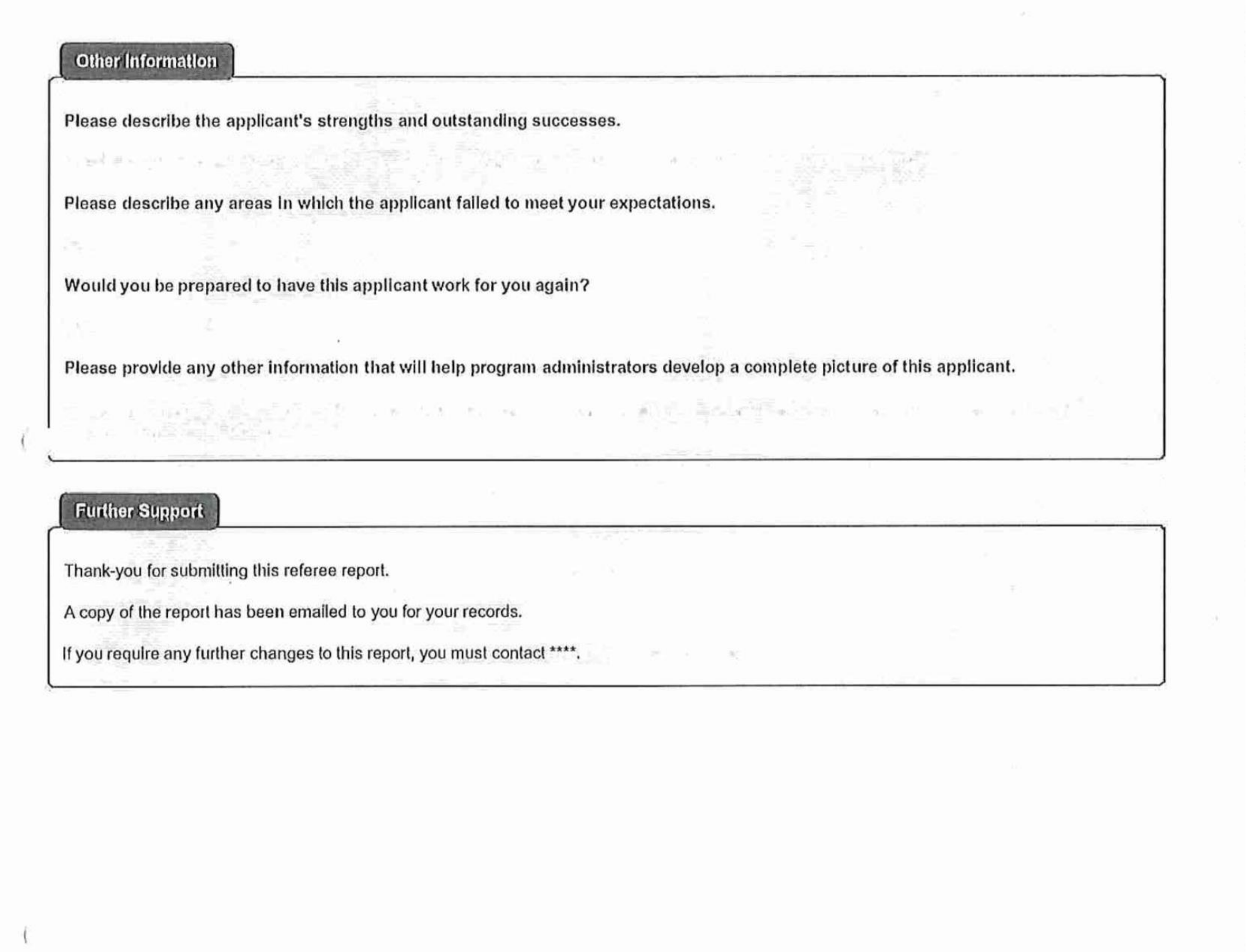
Performs above expected level

Time management skills
Organises and prioritises tasks in an effective manner.

Exceptional performance

Teamwork
Works with and contributes effectively within a team.

Exceptional performance



Other Information

Please describe the applicant's strengths and outstanding successes.

Please describe any areas in which the applicant failed to meet your expectations.

Would you be prepared to have this applicant work for you again?

Please provide any other information that will help program administrators develop a complete picture of this applicant.

Further Support

Thank-you for submitting this referee report.

A copy of the report has been emailed to you for your records.

If you require any further changes to this report, you must contact ****.

Appendix C 2015 Marking Guide

This appendix has been removed due to concerns about confidentiality and to protect the security of future selection processes.

Appendix D 2015 Advertisement training positions

Advertisement for **** positions 2015

1st Year ** Training.** This program is the entry level for ****. The program includes rotations in ****. Applicants who are already in their first year (or higher) of **** Training need to apply for a position via the **** Website.

The **** Training Program is run by a State-wide Network comprising supervisors representing each of the hospitals that offer **** Training. If you are successful in gaining a position, you will be required to rotate across all of these sites. For more information please contact ****.

If shortlisted for the selection process, you will be required to undertake several tasks which may include answering clinical scenario questions, possibly MCQs, and a communication station. Your CV and Covering Letter will be scored, as will your level of diligence and accuracy in completing the required documentation, and professionalism during the process. These will be scored in addition to the face-to-face interview, to enable us to develop a thorough assessment of your experience and capabilities. It also enables us to provide advice to those unsuccessful shortlisted candidates who request feedback.

In addition to your standard **Covering Letter** please ensure that you include a **reflection** on your biggest professional failure, be it clinical or non-clinical. Explore how you managed that situation and what you have learnt from it.

Appendix E 2016 Interview questions marking guide

This appendix has been removed due to concerns about confidentiality and to protect the security of future selection processes.

Appendix F 2016 Shortlisting marking

This appendix has been removed due to concerns about confidentiality and to protect the security of future selection processes.

Appendix G 2016 Application questions in lieu of cover letter

Questions to be included in the application form for **** Programs (in place of a cover letter);
Q1:
"To err is human", as the saying goes, implies that humans are by nature error prone, and when providing complex health care will inevitably make mistakes, and therefore we need to build and work within systems that minimise risks to both the patients and the professionals. Research in this area concludes that the vast majority of medical errors do not result from individual recklessness – it is mostly not a 'bad apple' problem.
Describe a clinical error that you were personally involved with and how, after reflection, this experience changed your clinical practice for the better.
(400 word limit)

Q2.
What strategies have you personally used to ensure that all clinical handover encounters, and requests for assistance, escalation or consultation, are successful? Specifically describe the principles underlying your strategies.
(250 word limit)

Q3.
Speaking up is a professional skill that significantly improves patient safety. However, speaking up in a traditionally hierarchical medical system is not easy for most people. Speaking up involves initiating a "crucial conversation", and some example situations include observed incompetence, poor teamwork, disrespect and other poor behaviours in team members, and there are others such as the scenario leader losing situational awareness. In fact, only about 10% of doctors readily and consistently initiate these crucial conversations, the other 90% rarely or never doing so (hence 'remaining silent', 'turning a blind eye', etc). This confirms that most people find speaking up difficult to do, in turn contributing to negative outcomes in healthcare.
Are the people who speak up crazy? Are they risking personal retaliation or career damage? Somewhat counter-intuitively, research shows that they achieve positive outcomes for patients, the hospital, and themselves. These clinicians have (or have learnt) the ability to deal with tough inter-personal challenges, and confront and resolve problems, and these skills are likely to be markers of other positive attributes.
Q: Taking into account the above, especially that most people find speaking up very challenging, please detail your personal reflection, especially what you learnt, on either –
Option 1: An experience where you felt unable or unwilling to 'speak up' or initiate a 'crucial conversation
OR
Option 2: An experience where you 'spoke up' or initiated a crucial conversation
(Choose Option 1 or 2. 400 word limit)
As an RMO at the **** you will inevitably have team members more junior than yourself, such as an intern or medical student. It is a reasonable expectation that you will participate in their learning and development, and their support and welfare, even though you are still a "junior doctor".

Q4.
What is the most important piece of advice that you would like to pass on to your juniors? Please explain why this is your best piece of advice.
(250 word limit)

Appendix H 2016 Applicant exit survey

**** **** Training Selection Centre Evaluation

You have now completed all the required tasks for the **** Selection Centre Process. We would value your feedback on the selection process.

All responses are anonymous and will not affect your selection.

Surveys will be analysed ONLY by an evaluator not involved in selection decisions, and after selection decisions have been made. **The interview panel members will not view the feedback sheets.**

Please answer honestly and provide as much information as you are able.

THANK YOU.

The Selection Process in 2015 has consisted of:
Curriculum Vitae
Short answer response questions submitted prior to today
Referee Reports
Multiple Choice Questions
Written Responses to Clinical Scenarios
Clinical Task (pathology Form)
Face to face interview

Considering the **overall selection process**:

	Strongly Disagree	Disagree	Agree	Strongly Agree
This process seemed to be an appropriate way to make decisions about selection into **** training				
I had the opportunity to demonstrate my skills and abilities				
This process required previous (post medical school) **** work experience to perform well				
This process was a fair assessment				
Please comment				

Task: Submit Curriculum Vitae

	Strongly Disagree	Disagree	Agree	Strongly Agree
This task seemed to be an appropriate way to make decisions about selection into **** training.				
I had the opportunity to demonstrate my skills and abilities.				
This task required previous (post medical school) **** work experience to perform well.				
This task was a fair assessment.				
Please comment				

Task: Short answer response questions (4) submitted prior to today

	Strongly Disagree	Disagree	Agree	Strongly Agree
This task seemed to be an appropriate way to make decisions about selection into **** training.				
I had the opportunity to demonstrate my skills and abilities.				
This task required previous (post medical school) **** work experience to perform well.				
This task was a fair assessment.				
Please comment				

Task: Referee Reports

	Strongly Disagree	Disagree	Agree	Strongly Agree
This task seemed to be an appropriate way to make decisions about selection into **** training.				
I had the opportunity to demonstrate my skills and abilities.				
This task required previous (post medical school) **** work experience to perform well.				
This task was a fair assessment.				
Please comment				

Task: Multiple Choice Questions

	Strongly Disagree	Disagree	Agree	Strongly Agree
This task seemed to be an appropriate way to make decisions about selection into **** training.				
I had the opportunity to demonstrate my skills and abilities.				
This task required previous (post medical school) **** work experience to perform well.				
This task was a fair assessment.				
Please comment				

Task: Written Responses to Clinical Scenarios

	Strongly Disagree	Disagree	Agree	Strongly Agree
This task seemed to be an appropriate way to make decisions about selection into **** training.				
I had the opportunity to demonstrate my skills and abilities.				
This task required previous (post medical school) **** work experience to perform well.				
This task was a fair assessment.				
Please comment				

Task: Clinical Task (Pathology Form)

	Strongly Disagree	Disagree	Agree	Strongly Agree
This task seemed to be an appropriate way to make decisions about selection into **** training.				
I had the opportunity to demonstrate my skills and abilities.				
This task required previous (post medical school) **** work experience to perform well.				
This task was a fair assessment.				
Please comment				

Task: Interview Panel

	Strongly Disagree	Disagree	Agree	Strongly Agree
This task seemed to be an appropriate way to make decisions about selection into **** training.				
I had the opportunity to demonstrate my skills and abilities.				
This task required previous (post medical school) **** work experience to perform well.				
This task was a fair assessment.				
Please comment				

Any additional comments you would like to make

Thank you for your feedback