






## Utility of selection methods for specialist medical training: A BEME (best evidence medical education) systematic review: BEME guide no. 45

Chris Roberts, Priya Khanna, Louise Rigby, Emma Bartle, Anthony Llewellyn, Julie Gustavs, Libby Newton, James P. Newcombe, Mark Davies, Jill Thistlethwaite & James Lynam


To cite this article: Chris Roberts, Priya Khanna, Louise Rigby, Emma Bartle, Anthony Llewellyn, Julie Gustavs, Libby Newton, James P. Newcombe, Mark Davies, Jill Thistlethwaite & James Lynam (2018) Utility of selection methods for specialist medical training: A BEME (best evidence medical education) systematic review: BEME guide no. 45, *Medical Teacher*, 40:1, 3-19, DOI: [10.1080/0142159X.2017.1367375](https://doi.org/10.1080/0142159X.2017.1367375)


To link to this article: <https://doi.org/10.1080/0142159X.2017.1367375>

 View supplementary material 



 Published online: 28 Aug 2017.

 Submit your article to this journal 

 Article views: 648

 View related articles 

 View Crossmark data 

 Citing articles: 1 View citing articles 

## Utility of selection methods for specialist medical training: A BEME (best evidence medical education) systematic review: BEME guide no. 45

Chris Roberts<sup>a</sup>, Priya Khanna<sup>b</sup>, Louise Rigby<sup>c</sup>, Emma Bartle<sup>d</sup>, Anthony Llewellyn<sup>e,f</sup>, Julie Gustavs<sup>b</sup>, Libby Newton<sup>b</sup>, James P. Newcombe<sup>g</sup>, Mark Davies<sup>h</sup>, Jill Thistlethwaite<sup>i</sup> and James Lynam<sup>j</sup>

<sup>a</sup>Primary Care and Medical Education, Sydney Medical School, University of Sydney, New South Wales, Australia; <sup>b</sup>The Royal Australasian College of Physicians, New South Wales, Australia; <sup>c</sup>Health Education and Training Institute, New South Wales, Australia; <sup>d</sup>School of Dentistry, University of Queensland, Queensland, Australia; <sup>e</sup>Hunter New England Local Health District, New Lambton, Australia; <sup>f</sup>Health Education and Training Institute, University of Newcastle, Newcastle Australia; <sup>g</sup>Royal North Shore Hospital, New South Wales, Australia; <sup>h</sup>Royal Brisbane and Women's Hospital, Queensland, Australia; <sup>i</sup>School of Communication, University of Technology Sydney, New South Wales, Australia; <sup>j</sup>Calvary Mater Newcastle, University of Newcastle, New South Wales, Australia

### ABSTRACT

**Background:** Selection into specialty training is a high-stakes and resource-intensive process. While substantial literature exists on selection into medical schools, and there are individual studies in postgraduate settings, there seems to be paucity of evidence concerning selection systems and the utility of selection tools in postgraduate training environments.

**Aim:** To explore, analyze and synthesize the evidence related to selection into postgraduate medical specialty training.

**Method:** Core bibliographic databases including PubMed; Ovid Medline; Embase, CINAHL; ERIC and PsycINFO were searched, and a total of 2640 abstracts were retrieved. After removing duplicates and screening against the inclusion criteria, 202 full papers were coded, of which 116 were included.

**Results:** Gaps in underlying selection frameworks were illuminated. Frameworks defined by locally derived selection criteria, and heavily weighed on academic parameters seem to be giving way to the evidencing of competency-based selection approaches in some settings.

Regarding selection tools, we found favorable psychometric evidence for multiple mini-interviews, situational judgment tests and clinical problem-solving tests, although the bulk of evidence was mostly limited to the United Kingdom. The evidence around the robustness of curriculum vitae, letters of recommendation and personal statements was equivocal. The findings on the predictors of past performance were limited to academic criteria with paucity of long-term evaluations. The evidence around nonacademic criteria was inadequate to make an informed judgment.

**Conclusions:** While much has been gained in understanding the utility of individual selection methods, though the evidence around many of them is equivocal, the underlying theoretical and conceptual frameworks for designing holistic and equitable selection systems are yet to be developed.

### Introduction

Specialty training programs aim to produce doctors who are capable of high quality, safe and independent practice. Selection of medical graduates into these programs is a high-stakes assessment process, which aims to predict the likelihood of applicants undertaking specialty training successfully and to identify those who are likely to perform poorly both in training and in future practice (Roberts and Togno 2011).

Selection processes are underpinned by two core aspects. First is a “predictive paradigm” where the intention is to predict who will be a competent doctor with expertise in the relevant specialty (Patterson and Ferguson 2010). In general, there is a lack of consensus in defining specific characteristics indicative of both a successful trainee and a doctor (Moore et al. 2015). While institutions such as the Royal College of Physicians and Surgeons of Canada and the Accreditation Council for Graduate Medical Education (US) have developed frameworks of standards that provide an overarching scaffold of defined domains of competence, (Frank and Danoff 2007), there is little research on the extent to which these frameworks have informed the

### Practice points

- Locally-defined selection systems found to be subjective and heavily weighed on academic parameters.
- Selection systems using competency-based approaches are gradually evolving, though the evidence is contextualized. Multiple selection tools in such systems had favorable evidence.
- Predictive validity mostly limited to academic criteria with methodological issues and paucity of long-term evaluations.

selection process. Recently, job analysis techniques have been used to assist training institutions in identifying core and specialty-specific academic and nonacademic skills and frame these as assessable competencies for selection into postgraduate training (Patterson et al. 2008); however, the evidence is limited.

The second paradigm underlying selection is as a high stakes assessment; therefore, principles underlying any good assessment should be considered when designing

frameworks and methods (Van Der Vleuten 1996; Prideaux et al. 2011). As in any assessment, Van der Vleuten's utility index has been used widely to capture psychometric robustness of selection methods (Thomas et al. 2012). Utility is defined as a multiplicative function of reliability, validity, educational impact, acceptability, feasibility and cost-effectiveness (Van Der Vleuten 1996). Reliability refers to the reproducibility or consistency of scores from one assessment to another. It is best measured by a generalizability coefficient, which estimates multiple sources of error and provides a method that is generalizable to virtually any setting (Cook and Beckman 2006).

The trustworthiness of assessments is a question of validity. In selection, predictive validity refers to how well a tool identifies applicants who will display desired attributes upon graduation and throughout their professional practice (Cameron et al. 2017). Test scores and grades alone are insufficient to select applicants as they tap only a narrow band of the complex and multidimensional role of a specialist doctor (Hamdy et al. 2006). Within undergraduate medical training, there have been several efforts to examine the predictive attributes of both academic and nonacademic factors influencing success (Eva et al. 2009; Patterson, Knight, et al. 2016). However, fewer studies have focused on predictors of success in postgraduate training and, of these, the majority are centered around cognitive or academic factors (Ferguson et al. 2002; Tolan et al. 2010). Noncognitive attributes, which might predict success in specialty training include integrity, reliability, diligence, trustworthiness, commitment, respect and empathy and interpersonal skills such as communication and team work. Evidence is limited owing to difficulty in obtaining quantifiable and reliable data (Bernstein et al. 2003; Egol et al. 2011; Schaverien 2016).

In recent years, the concept of validity has been extended to include social validity that captures fairness of selection procedures and outcomes as underpinned by organizational justice theory (Colquitt et al. 2001). Extending the concept of social validity beyond the applicant and the organization, Patterson et al. (2012) refer to the concept of "political validity," which includes sociopolitical and other stakeholder groups that may influence the design and development of selection systems.

### **Selection methods**

Globally, there are marked variations in selection procedures for specialty training across various countries. In the United States, for instance, selection relies on a national match system for selecting applicants to the program. Locally determined selection processes are supported by a range of data including past academic records, scores in standardized licensing examinations, curriculum vitae, personal statements, referees' reports, Dean's letters and letters of recommendations (McGaghie et al. 2011; Krauss et al. 2015; Katsufakis et al. 2016; Sklar 2016). Although in Canada, selection also relies on locally defined criteria; there is an increasing move towards aligning criteria with competency-based medical education principles.

Elsewhere, the United Kingdom and Australia have made systematic efforts in developing robust and defensible selection procedures using a wide range of written and

observed formats. The selection methods may be either low fidelity (such as written or video scenario-based tests) or high-fidelity methods (such as simulations that replicate authentic job-related tasks). An evidence base is emerging on several selection methods, including: multiple mini-interviews, situational judgment tests, clinical problem-solving tests, simulations and selection/assessment centers (Patterson, Carr, et al. 2009; Roberts and Togno 2011; Patterson, Rowett, et al. 2016).

Multiple mini-interviews (MMIs) have been used to assess noncognitive characteristics of entry-level medical students and more recently postgraduate trainees. They are based on the objective-structured clinical examination (OSCE) format, comprising short interview stations, each with different examiners. At each station, the applicant is presented with a question, hypothetical scenario, or task (Eva et al. 2004; Roberts et al. 2008). Currently, MMIs are being used for postgraduate training selection internationally including in the United Kingdom, Canada (Dore et al. 2010) and Australia (Roberts et al. 2014). Pilot implementations have been undertaken in Japan (Yoshimura et al. 2015), the Middle East (Ahmed et al. 2014) and Pakistan (Andrades et al. 2014).

Situational judgment tests (SJT) are used to assess applicants' noncognitive characteristics by presenting them with hypothetical written or video-based scenarios of a situation they are likely to encounter in job roles. Applicants are required to choose the most appropriate responses or to rank the responses in the order they feel reflects the most appropriate course of action. SJTs have been regarded as an approach to measurement rather than a single style of assessment, as the scenario content, response instructions and format vary widely across settings and specialties (Patterson, Zibarras, et al. 2016). They have been introduced into the selection processes of several medical specialties within the United Kingdom and into the Australian general practice training (Patterson, Zibarras, et al. 2016).

Clinical problem-solving tests (CPST) are based on multiple-choice question-formats. The CPST presents clinical scenarios for applicants to apply their clinical knowledge in order to solve a problem reflecting, for example, a diagnostic process or to develop a patient management strategy (Patterson, Baron, et al. 2016). Currently, the CPST is being used as one of the assessments for selection into a range of specialties in the United Kingdom, where it is usually combined with an assessment of noncognitive factors such as the SJT.

Selection/assessment centers allow an applicant to participate in multiple processes comprising a number of job-related assessments such as written exercises, interviews, group discussions and simulations. While selection or assessment centers have been used in several occupational groups, its use in medical selection system is relatively new and was initiated in the national training selection processes in the United Kingdom and in Australia (Gale et al. 2010; Roberts and Togno 2011; Pashayan et al. 2016).

Given that selection processes into specialty training involves high-stakes decisions, it is important for training institutions to adopt an evidence-based approach in designing, implementing and improving criteria and methods. With this aim in mind, we undertook the current review. While there is a substantial literature focusing on selection into medical school, we were unable to find a

comprehensive review on the criteria and methods of selection specifically into postgraduate training.

## Review aims

The goal of this review was to explore, analyze and synthesize the evidence related to selection into postgraduate medical specialty training, through the following research questions.

1. What are the underlying frameworks, principles and methods of selection into postgraduate medical specialty training?
2. How effective are the existing methods and criteria in terms of validity, reliability, feasibility, acceptability, cost-effectiveness and other indicators of a good assessment?
3. What are the predictors of success in subsequent performance?

## Review method

### Pilot phase

The Topic Review Group (TRG) included members from a diverse range of disciplines within postgraduate medical education and research. Prior to the full systematic review, a partial pilot review for the articles published between 2010 and 2013 (336 relevant abstracts) was conducted to test the proposed review protocol. The pilot review helped in establishing the search strategy, and inclusion/exclusion criteria and trialing of the review coding forms. It also helped in the refinement and sensitivity of the search syntax to enhance its relevance and wider postgraduate specialty coverage.

### Study selection

Acceptable study designs for the main review based on the study criteria (Table 1) included prospective and retrospective studies, cross-sectional and longitudinal studies, as well as systematic literature reviews. Acceptable data included qualitative, quantitative, mixed or multiple data using relevant data collection methods such as surveys, observations, interviews or focus groups. Empirical data collection focused on the components of the utility of any assessment (Van Der Vleuten 1996).

### Search strategy

The search strategy was aligned with the recommendations of Haig and Dozier (2003) who assert that core principle databases should be consulted and secondary databases should be employed according to the nature of search topic. The electronic database PubMed was searched using

the search syntax. Other core bibliographic databases, such as Ovid Medline, Embase, CINAHL; ERIC and PsycINFO (using EBSCO), were also searched along with hand-searching key journals, and new abstracts were reconciled with the ones retrieved using PubMed. We retrieved a total of 2,640 abstracts, which were imported into EndNote X5. After removing the duplicates and screening against our inclusion criteria, a total of 202 full papers were retrieved and coded, of which 116 were included (including the pilot study articles) (Figure 1). All titles/abstracts were entered into a dedicated EndNote library.

In this study, we defined “trainee” or “resident” as a medical graduate who intended to commence further training in a postgraduate training program in various specialties related to direct patient care. The term, “standardized methods or tests” in this review refers to assessment methods in which the questions, conditions of test administration, scoring procedures and interpretations of results are consistent (Ahmed et al. 2017).

### Coding process

The standard BEME coding sheet was modified in light of the review questions to extract relevant data. Coding sheet can be viewed in the Supplementary files. Papers ( $n=116$ ) were divided among five pairs from the TRG with each pairing independently reviewing the full text using the agreed coding sheet. The pair then discussed and negotiated any divergent opinions and developed a consensus coding sheet. If a consensus could not be reached, a third reviewer was approached for resolution.

### Data analysis

A spreadsheet organized summaries of completed coding forms, informing the descriptive breakdown of the number of papers by strength of evidence and overall impressions.

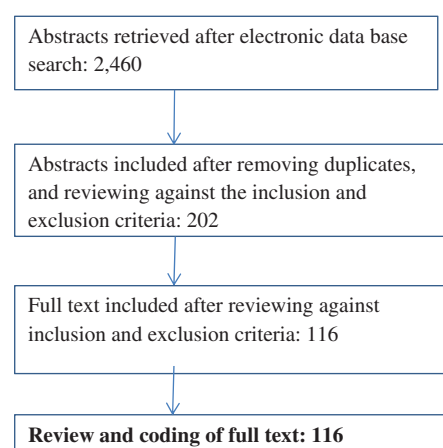


Figure 1. Flowchart of literature search and paper selection.

Table 1. Inclusion and exclusion study criteria.

Inclusion criteria	Exclusion criteria
<ul style="list-style-type: none"> <li>• Work-based postgraduate specialty training</li> <li>• Clinical discipline</li> <li>• Focus of paper on selection into specialty training program</li> <li>• Empirical data on selection</li> <li>• Published between 1 January 2000 and 31 May 2016</li> <li>• In English</li> </ul>	<ul style="list-style-type: none"> <li>• Medical school selection</li> <li>• Health professions other than medicine</li> <li>• Focus of paper on aspects unrelated to selection such as career choice</li> <li>• No empirical data</li> <li>• Published before 1 January 2000</li> <li>• Not in English</li> </ul>

We rated papers as high quality if they ranged between “results are unequivocal” to “conclusions can be somewhat drawn” in terms of strength of findings; and they were rated as “excellent”, “good” and “acceptable” in terms of overall impression. A total of 89 articles were rated as high quality.

## Review findings

We synthesized the findings in line with our three research questions.

### **1. What are the underlying frameworks, principles and methods of selection?**

We defined assessment frameworks and principles as the conceptual and theoretical underpinnings that inform the development of selection systems, processes, criteria and methods. In terms of categorizing various selection models, we found two major types of selection frameworks: those based on locally defined selection criteria, and those based on well-defined criteria with multiple selection methods.

In some countries such as the United States, the selection systems are based on locally defined selection criteria that are subjective, at the discretion of the specific program directors or selectors of the programs locally and rely more on past academic attainment. In contrast, frameworks as used in, for example the United Kingdom and Australia, involve multiple methods of selection with more globally defined selection criteria.

#### **Selection systems based on locally defined selection criteria**

In the United States, an applicant may enter residency (specialist) training programs after successful completion of medical school and having passed the first two steps of a three-step United States Medical Licensing Examination (USMLE) examination. Specialty training starts in the first year after graduation (also known as the intern year). Applicants submit an online application and supporting documents using the Electronic Residency Application Service (ERAS). Interviews are undertaken for the chosen program, while the applicants are still medical students, after which applicants and program directors each rank their respective parties (rank-order list). The National Resident Matching Program (NRMP) uses a uniform residency application and administers the match of the rank-order lists using a computer algorithm, and on the “match day” applicants are notified of the program they have been matched with (Sbicca et al. 2010).

While the post-interview rankings are based on predetermined formulas and uniform criteria such as the USMLE scores, there are several other subjective factors influencing selection such as subjectivity in the interview scoring, letters of recommendations, prior research and clinical experience. The relative importance of these criteria is at the discretion of the specific program directors of the programs locally.

#### **Selectors’ and applicants’ perceptions of locally defined selection criteria.**

Six articles explored program directors’

perceptions about the selection criteria and their relative importance in selecting residents through anonymous surveys. Interview scores and USMLE Step 1 and 2 were the most valued factors in the final selection of the applicants, followed by letters of recommendations (Makdissi et al. 2011; Al Khalili et al. 2014). Crane and Ferraro (2000) reported specialty-specific (Emergency Medicine) rotation grade, clinical grades and interview to be the most important, whereas USMLE scores and recommendations were found to be moderately important selection criteria. While Makdissi et al. (2011) found prior research experience and publications in general surgery as the least important screening factors; Melendez et al. (2008) reported that basic science and clinical research by applicants was always considered for their general surgery training programs.

Two studies explored program directors’ views of plastic surgery training programs in the United States where some applicants are directly from medical school (integrated path) and some have completed other specialty training (such as general surgery and urology). Janis and Hatef (2008) investigated program directors’ views on selection criteria in the integrated training program, whereas Nguyen and Janis (2012) surveyed program directors of the independent training pathway. Program directors of both programs perceived letters of recommendation and interviews to be among the most important factor for selection. Those in the integrated pathway also preferred subinternship rotation performance, whereas those in the independent pathway emphasized USMLE Step 1 scores.

#### **Factors influencing selection system and outcomes.**

Seven studies reported data on correlates of successful match outcomes, by either surveying applicants or retrospective review of the ERAS documents. The USMLE Step 1 scores and successful acceptance into the Alpha Omega Alpha Honor Medical Society (AOA) were among the most common factors to be positively correlated with successful matching or with number of interview invitations (Baldwin et al. 2009; Rogers et al. 2009; Fraser et al. 2011; Stratman and Ness 2011; Maverakis et al. 2012). The AOA is a professional medical organization that recognizes and advocates for excellence in scholarship and the highest ideals in medicine.

While the authorship of one or more peer-reviewed publications was found to correlate with favorable match outcomes (Rogers et al. 2009; Fraser et al. 2011; Stratman and Ness 2011), the quality of publication as determined by the journal impact factor did not appear to have positive impact on the outcome (Stratman and Ness 2011; Maverakis et al. 2012). The use of authorship may be a possible source of applicant self-inflation in the match process, as Maverakis et al. (2012) reported that successful applicants listed multiple in-preparation manuscripts, the majority of which were subsequently found to be unpublished. Conflicting evidence was reported for class rank. While Rogers et al. (2009) reported high class-rank to be significantly associated with the number of interview invitations, Baldwin et al. (2009) found class rank and medical school grades to have little effect on the match success. Other factors associated with successful matching included letters of recommendation (Fraser et al. 2011; Stratman and Ness 2011), away rotations in the area of chosen specialty



(Baldwin et al. 2009) and applicants' satisfaction with the match process itself (Lansford et al. 2004). Robinson et al. (2013) assessed the Residency Training Coordinator (RTC) role in predicting psychiatry resident applicants' success in obtaining a residency position. RTCs are responsible for organizing and disseminating necessary application materials from applicants to facilitate the selection process. The authors found that all the applicants who successfully matched in the psychiatry residency program had received higher scores from the RTCs, and concluded that RTCs can provide an important perspective on residency applicants' attentiveness, communication, attitude and professionalism.

A few studies examined the integrity of the matching program and whether the process was biased against particular cohorts of applicants. Sbicca et al. (2010) surveyed Stanford dermatology residency applicants, residents and program directors revealing some NRMP policy violations as well as ethical infraction by some program directors during their communications with applicants. Despite the underrepresentation of women in orthopedics, Scherl et al. (2001) found no evidence of gender bias against women applicants in the initial review of application for residency. In another study, Chew et al. (2005) examined the utility of a computer software (spreadsheet) designed to address scoring variability in the match-list for radiology residency selection and found it to be fair, objective and efficient.

### ***Selection frameworks based on well-defined criteria with multiple methods***

In the United Kingdom, the principles of organizational psychology have been used to identify and develop selection criteria and methods by identifying core and specialty-specific competencies. Using the tenets of job analysis, Patterson et al. (2008) undertook three multisource, multi-method independent studies to explore core and specific competencies in anesthesia, obstetrics and gynecology and pediatrics. The outcome comprised 14 general competency domains common to all specialties. This study was replicated by Patterson, Tavabie et al. (2013) to explore competencies for general practice training which resulted in 11 competency domains, of which empathy and perspective-taking, communication skills, clinical knowledge and expertise and professional integrity were rated as the most important domains. Patterson et al. (2014) extended the competency model approach to examine specific knowledge, skills and attributes associated with the roles of assessors and simulations in the GP selection centers in the United Kingdom. In examining applicants' reactions following the shortlisting stage and after the selection center (interview) stage, Patterson et al. (2011) reported that, of all the selection methods, the simulated patient consultation (high-fidelity) undertaken at the selection center was rated as most job-relevant and therefore most valid.

In summary, selection systems based on criteria defined by the local program directors or selectors seem to place more emphasis on applicants' past academic achievement although lack of studies make data comparison and generalization difficult. By contrast, selection frameworks based on well-defined selection criteria and using the principles of organizational psychology tend to be more objective, and seem to go beyond the discretion of selectors of the individual training program. The number of studies

investigating these frameworks was low, and limited mostly to UK speciality selection systems.

### ***How effective are the existing methods and criteria in terms of validity, reliability, feasibility, acceptability, cost-effectiveness and other indicators of a good assessment?***

Fifty studies were related to the following main methods of selection into specialist training: traditional interviews and multiple mini-interviews (MMI); situational judgment test (SJT); clinical problem-solving test (CPST) and selection centers/assessment centers. We also found that in several specialties (especially in North America), selection is heavily reliant on selection criteria such as letters of recommendation (LOR), licensing examinations, specialty-specific aptitude tests, and other academic and nonacademic criteria.

#### ***Interviews***

***Range of evidence.*** Of 20 studies with data on the utility of interviews, 11 were related to MMIs. Four studies involved retrospective analysis of data, one was a systematic review, and the rest were based on a prospective study design. Twelve studies were based on a quantitative approach, and seven used mixed methods. Nine considered an aspect of interviews, eight of the studies had a main focus on the multiple mini-interviews (MMI), and three included a comparison between traditional interviews and MMIs. The number of applicants involved ranged from 14 (Andrades et al. 2014) to 1382 (Roberts et al. 2014).

***Number of stations.*** For the MMI studies, the number of stations ranged from four (Soares III et al. 2015) to twelve (Hofmeister et al. 2009) and the average number of stations was between seven and eight.

***Reliability.*** The reliability of the interview process was examined by eleven of the studies. In one study, the intra-class correlation coefficient (ICC) of MMIs was used as a measure of inter-rater reliability of interviewers, and ranged between 0.24 to 0.98 although the majority were above 0.8 (Campagna-Vaillancourt et al. 2014). Generally, the reliability of the multiple-mini interviews (MMI) (derived from generalizability theory) was considered acceptable (ranging from 0.55 to 0.72) (Dore et al. 2010). On comparing behavioral and situational type of MMI formats, Yoshimura et al. (2015) found a seven-station MMI, in either type gave an inter-rater reliability of more than 0.80. Elsewhere the reliability of a six-station MMI of the behavioral type had a generalizability co-efficient of 0.76 (Roberts et al. 2014).

The overall reliability of structured interviews was reported as high. In one study (Bandiera and Regehr 2004), the reliability (internal consistency) as determined by Cronbach's alpha for four interviews was 0.83. Inter-rater reliabilities within interview pairs ranged from 0.37 to 0.69, whereas inter-rater reliabilities between interviewers from different interviews ranged from -0.13 to 0.69. The authors suggested that interviewers based their scores on an overall global impression despite interviewer training. A Danish study (Isaksen et al. 2013) on selection into family medicine used semistructured interviews that combined

individualized elements from the applications with standardized behavior-based questions. There was high internal reliability (Cronbach's  $\alpha = 0.97$ ) for the first selection round using only standardized behavioral questions based on key roles; and 0.90 for the second selection round using standardized behavioral questions combined with the themes from applicants' form. However, the generalizability coefficient of the first round was 0.74 and 0.40 for the second round suggesting further development of the tool was required. These reliability results are not dissimilar to those found in the undergraduate selection setting (Eva and Macala 2014).

**Validity.** Within Australian GP training selection, the MMI had reasonable construct and concurrent validity (Roberts et al. 2014). Performance in a six-station MMI predicted three end-of-training assessments; a knowledge test ( $r = 0.12$ ), key features test ( $r = 0.24$ ) and an OSCE ( $r = 0.46$ ). Prediction when combined with the SJT, improved for the key features and OSCE, but not the knowledge tests. This suggested that MMI and SJT were complementary, as they both explained incremental variance over each other for end-of-training assessments (Patterson, Rowett, et al. 2016).

Of those studies that investigated the predictive value of traditional interview scores for success in subsequent performance, the majority reported positive findings. Alterman et al. (2011) concluded that the interview scores (traditional format) of general surgery residency applicants could predict successful completion of training. However, the results, showing an odds ratio of 118.27 with a very wide 95% confidence interval, (3.757–9435.405) for a small sample size ( $n = 101$ ) were met with skepticism because of the lack of accuracy in the estimate. Another study on general surgery residency found that the personal characteristics and letters of reference were predictive of subsequent clinical performance ratings on core competencies (ranging from  $r = 0.15$ – $0.45$ ) (Brothers and Wetherholt 2007).

While not causal, this same study (Brothers and Wetherholt 2007) reported the correlation of a combination of two interviewer-based tools and the final match list. One was a "personal characteristics tool" that captured the impressions of the faculty interviewer of the candidate's attitude, motivation, integrity, interpersonal relationships and response to specific life challenges. The other recorded the interviewers' assessment of the applicants' letters of reference. Taken together, these predicted the final match list ( $r = -0.76$ ), as favorable correlations are negative with greater selection scores correlating with lower ordinal rank number.

Two studies around the predictive value of applicants' rank generated post the interview process appeared contradictory. Olawaiye et al. (2006), found the rank list, which had been generated using structured interviews for the NRMP, was significantly correlated with first-year clinical performance ( $r = 0.60$ ). In a retrospective review, Adusumilli et al. (2000) found no correlation between the faculty generated rank number and residents' performance in rotation evaluations or board examinations.

In the Australasian context, Oldfield et al. (2013) found positive but small associations between semistructured multi-station interview scores and formative assessments (miniclinical evaluation exercise and a clinical examination),

as well as with the summative clinical examination for surgical trainees. Lillis (2010) examined interview scores for GP training applicants and reported moderately strong correlations with the summative written and clinical examination scores. On examining the association between selection factors and subsequent performance among international medical graduates applying for psychiatry residency, Shiroma and Alarcon (2010) reported a negative correlation ( $r = -0.20$ ) for an in-training written examination, but positive with a work-based assessment ( $r = 0.38$ ).

Two other studies also reported contradictory findings in terms of selection interviews predicting residency performance. Bell et al. (2002) and Khongphatthanayothin et al. (2002) found no correlation between interview scores and subsequent evaluation of resident performance in pediatric and obstetrics and gynecology, respectively.

**Acceptability.** Acceptability to applicants and faculty is a core concern of any admissions process. Overall, we found favorable evidence: MMIs were considered fair by applicants and it improved the assessors' judgment (Isaksen et al. 2013); it was considered more accurate by applicants and assessors alike (Dore et al. 2010); and that it was free from gender and cultural bias (Hofmeister et al. 2009). Not all the studies were supportive for MMI's acceptability. One study reported the presence of a MMI might impact their decision to interview at that program (Hopson et al. 2014). In another, US emergency medicine residency applicants preferred traditional interviews over MMIs. This was due to multiple factors, principally lack of familiarity with the MMI, inability to form a personal connection with the interviewer and difficulty perceiving fit with the program (Soares III et al. 2015).

**Feasibility.** The feasibility of interviews and MMIs was reported with mixed results. In one study, four out of a total of eight interviewers considered MMIs to be feasible (Andrades et al. 2014). Others highlighted that MMIs can present additional work to set up in year one, but less in future years (Campagna-Vaillancourt et al. 2014). These equivocal findings in the postgraduate sector resonate with the observations in the systematic review into student selection (Pau et al. 2013) that MMIs did not require more examiners when compared to the panel interview, did not cost more, interviews could be completed over a short period of time and could be a positive experience for both interviewers and applicants.

#### **Situational judgment tests (SJT)**

**Range of Evidence.** Eleven studies focusing on situational judgment tests (SJTs) were reviewed. Of these, six were longitudinal quantitative studies, three were cross-sectional quantitative studies, one was a systematic review, and two were nonsystematic reviews.

**Reliability.** Overall, SJTs appeared to demonstrate high reliability and validity especially within the general practice setting in the United Kingdom and Australia. In a pilot testing in the UK GP setting, internal reliability was reported in the range of  $r = 0.80$ – $0.83$  (Patterson, Baron, et al. 2009). The reliability of the SJT used in GP selection in Australia was reported to be 0.91 (Patterson, Rowett, et al. 2016). However, the internal reliability of the SJT used in a pilot

for selection into Dutch GP training was more modest at 0.55 (Vermeulen et al. 2014). This was perceived to be due to the limited number of situations that were tested or contextual issues related to the Netherlands. It should be noted that the marking system was also very different from the UK SJT marking system.

**Validity.** In pilot testing for selection processes into UK GP training, the SJT correlated with the clinical problem-solving test (CPST), varying from  $r=0.39$  (Patterson, Baron, et al. 2009) to  $r=0.53$  (Patterson, Lievens et al. 2013). In core medical training in the United Kingdom, the correlation ranged from  $r=0.45$ – $0.53$  (Patterson, Carr, et al. 2009).

There was a correlation with structured application forms in the UK GP training setting  $r=0.41$  (Patterson, Baron, et al. 2009). In a Dutch GP setting, the SJT correlated with a knowledge test ( $r=0.14$ ) and a structured interview ( $r=0.34$ ) (Vermeulen et al. 2014). In an Australian GP setting, the SJT correlated with MMI performance ( $r=0.39$ ) (Patterson, Rowett, et al. 2016).

There was some variation in SJTs in correlation with performance in work-based simulations, varying from  $r=0.40$  (Koczwara et al. 2012) to  $r=0.72$  (Ahmed et al. 2012a) within a selection center setting for selection into general practice in the United Kingdom. In shortlisting into core medical training in the United Kingdom, the SJT correlated ( $r=0.53$ ) with the structured interview outcomes.

Two studies reported the SJT predicting end-of-training performance. In the United Kingdom, in a sample of  $n=2292$ , the SJT predicted an end-of-training applied knowledge test ( $r=0.43$  – corrected to 0.69 for range restriction) and an end-of-training objective structured clinical examination (OSCE) ( $r=0.43$ , corrected to 0.57 for range restriction). The SJT also correlated with a three station simulation exercise undertaken within a selection center (Patterson, Lievens, et al. 2013). These findings were reproduced within the Australian GP setting. The SJT predicted end-of-training applied knowledge test ( $r=0.14$ ), a key feature problem test ( $r=0.24$ ) and an end-of-training OSCE ( $r=0.44$ ). However, these coefficients were not corrected for range restriction (Patterson, Rowett, et al. 2016).

Three studies also reported on the incremental contribution to predictive validity that the SJT has in combination with other instruments. In the UK medical training setting, the combination of the CPST and the SJT predicted the final interview scores, with the SJT adding an additional 15% of the variance, to increase the predictive validity of the combined machine-marked tests (Patterson, Carr, et al. 2009; Koczwara et al. 2012). Furthermore, in the Australian GP setting, both the SJT and MMI contributed to incremental validity over each other, the SJT greater in predicting knowledge tests and the MMI in predicting OSCE and written key feature problem tests (Patterson, Rowett, et al. 2016).

**Acceptability and feasibility.** A systematic review of the SJT reported that there is acceptability evidence in the organizational psychology literature for the SJT (Patterson, Knight, et al. 2016). Within postgraduate medical education literature, the acceptability of the SJT is equivocal. In a large sample ( $n=2947$ ), there was a good agreement among respondents (>60%) that the content of SJTs was

clearly relevant to GP training and appropriate for the entry level they were applying for. However, only a third agreed that the test gave them sufficient opportunity to indicate their ability for training and that the test would help selectors differentiate between applicants (Koczwara et al. 2012). In a qualitative study focusing on the social validity of selection processes in the Australian GP setting, although overall rating for the combination of the MMI and the SJT was positive, there were concerns about the acceptability of the SJT by a small minority of the sample (18%) (Burgess et al. 2014).

While none of the papers reported feasibility as an outcome directly, the assumption appears safe that the SJT is feasible, as it has been implemented and evaluated in at least three different countries, and within two disciplines. No cost-effectiveness data have been published within the medical education literature.

### **Clinical problem-solving tests (CPST)**

**Range of evidence.** Three longitudinal studies focused on the clinical problem-solving test using differing datasets from the same UK GP selection setting. A fourth paper reported a cross-sectional study from UK medical training with comparative data for general practice. All these papers were from the same research team.

**Reliability.** As with all reasonably long written tests, the CPST has good internal reliability ( $r=0.85$ – $0.89$ ) (Patterson, Carr, et al. 2009), as has been found in studies of in-training knowledge tests studies at the undergraduate level (e.g. MCAT correlating with USMLE Step 1 scores, or MCAT correlating with MCC in Canada) and might be expected to correlate with end-of-training knowledge tests (Prideaux et al. 2011).

**Validity.** Concerns with the construct validity of CPST have been raised. There was no firm evidence that the CPST validly tests problem-solving skills rather than knowledge (Patterson, Baron, et al. 2009; Crossingham et al. 2011). In pilot testing for selection processes into UK GP training, the CPST correlated with the SJT varying in the range of  $r=0.39$  to  $r=0.53$  (Patterson, Baron, et al. 2009; Patterson, Lievens, et al. 2013). In core medical training in the United Kingdom, the correlation ranged from  $r=0.45$ – $0.53$  (Patterson, Carr, et al. 2009). In the UK medical training setting, the CPST correlated with the final interview scores ( $r=0.34$ ) (Patterson, Carr, et al. 2009). However, there was a greater predictive validity of the combination of the CPST and the SJT. The CPST scores also correlated with a cognitive ability test consisting of verbal, numerical and visual-spatial ability ( $r=0.41$ ); and with a test of nonverbal ability ( $r=0.36$ ) and an overall assessment center score ( $r=0.38$ ) (Koczwara et al. 2012). The acceptability of CPST from the applicants' perspective was high across issues of relevance, fairness, opportunity to demonstrate ability and differentiation between applicants. The cost of the CPST was estimated to be \$30 (USD) for each applicant (Patterson, Baron, et al. 2009).

**Specialty-specific tests.** Four studies investigated the assessment of technical skills. Carroll et al. (2009) in plastic surgery, and Gallagher et al. (2008) in general surgery



investigated the usefulness of a previously validated Objective Structured Assessment of Technical Skills (OSATS) as a part of a selection process for higher surgical training. Carroll et al. (2009) noted that those selected into higher training performed 2.2 times better on average in a six station OSATS than those who were not selected. Gallagher et al. (2008) reported a strong relationship between performance in a 10 station OSATS in relation to overall performance in the program ( $r = 0.76$ ).

A retrospective analysis of the use of a specialty-specific written examination for selection into general surgery by Farkas et al. (2012) found the assessment to correlate more closely (than the licensing examination, USMLE) with an in-service examination undertaken during the first year of the training program. The authors acknowledged insufficiency of data to report reliability for the examination. Moore et al. (2015) found that an aptitude test (with a component to assess attitudes) designed for otolaryngology residency selection predicted performance during training.

**Selection centers.** Ten papers reported selection in the context of an assessment or selection center. One article (Patterson et al. 2014) was a qualitative study exploring competency models to improve uniformity and calibration of the overall process. Two articles described the Australian GP selection center process, two quantitatively (Roberts et al. 2014; Patterson, Rowett, et al. 2016) and one qualitatively (Burgess et al. 2014), two describing a selection center approach into anesthetics training (Gale et al. 2010; Roberts et al. 2013), three describing the UK GP selection center approach (Mitchison 2009; Patterson, Baron, et al. 2009; Patterson, Lievens, et al. 2013) and a systematic review (Patterson, Knight, et al. 2016).

In the UK GP setting, selection center refers to three job-relevant simulations (patient consultation, group and written simulation exercises) targeting both clinical and nonclinical attributes. In the UK GP setting, the selection center scores significantly correlated with the CPST ( $r = 0.30$ ) and the SJT ( $r = 0.46$ ). The selection center was predictive of supervisor rating after 1 year, which were used as a proxy for job performance ( $r = 0.30$ , corrected to 0.50 for restricted range) (Patterson, Baron, et al. 2009). The score for overall performance at selection achieved statistically significant correlation with examination performance ( $r = 0.49$ ) for the applied knowledge test and  $r = 0.53$  for the clinical skills assessment (Davison et al. 2006; Ahmed et al. 2012b). In the Australian setting, neither Roberts et al. (2014) nor Patterson, Knight et al. (2016) reported the composite selection center score. The principal concerns within the papers describing selection centers refer to assessment frameworks, acceptability, and feasibility, particularly around cost-effectiveness.

### Letters of recommendation (LOR)

**Range of evidence.** Four studies reported data on standardized letters of recommendation (SLORs), two of which were retrospective analyses (Love et al. 2013; Beskind et al. 2014), the other two were experimental studies conducted by the same research group (Prager et al. 2012; Perkins et al. 2013). Five retrospective studies examined the predictive value of letters of recommendation along with other selection criteria (Boyse et al. 2002; Khongphatthanayothin

et al. 2002; Hayden et al. 2005; Brothers and Wetherholt 2007; Oldfield et al. 2013).

**Reliability.** We found that concerns around low reliability with the traditionally used narrative letters of recommendation (NLORs) led to the development of SLORs. Two experimental studies looking at SLOR found higher interrater reliability of SLORs compared with NLORs (Prager et al. 2012; Perkins et al. 2013). However, these findings were contradicted by results from the two retrospective studies into SLORs, which showed inter-rater reliability was influenced by the experience of the reference writer (Love et al. 2013; Beskind et al. 2014).

**Validity.** Validity was addressed in two retrospective studies on SLORs, indicating the potential for grade inflation on SLORs, limiting their ability to discriminate between applicants (Love et al. 2013; Beskind et al. 2014). In terms of predictive validity of LORs, the evidence seems to be inconclusive. While evidence suggested the predictive value of LORs in subsequent clinical performance (Hayden et al. 2005; Brothers and Wetherholt 2007; Oldfield et al. 2013), Boyse et al. (2002) found no predictive value of a Dean's letter or superior letters of recommendations for future performance. Khongphatthanayothin et al. (2002), although reporting no predictive association between letters of recommendation and in-training examination, found a weak association with faculty clinical evaluations.

**Feasibility and acceptability.** Prager et al. (2012) and Perkins et al. (2013) demonstrated that SLOR templates were reasonably easily designed and, once implemented, are likely to be sustained as a process (over NLORs) as they are more time-efficient in terms of reviewers processing information and rating applicants.

**Personal statements and curriculum vitae (CV).** Only one study in our review was related to the quality and utility of personal statements (Max et al. 2010). Structural analysis and program directors' perceptions of de-identified personal statements revealed good inter-rater reliability for features of essays and common features written by the applicants within personal statements. However, the quality of the statements was perceived as less original and compelling and, when using the statements to differentiate between applicants, only a fraction of program directors found them to be "very important".

In regards to CVs, only one study in our review examined the validity of a CV in totality and reported negative associations between CV and subsequent formative and summative performance indicators among trainees in general surgery (Oldfield et al. 2013).

To summarize, we found favorable evidence on the reliability, construct, concurrent and predictive validity of interviews especially MMIs. The data on acceptability and feasibility of MMIs appeared to be mixed, and data around cost-effectiveness was limited.

In relation to the SJTs, generally the internal reliability of the tool was reported as high, and predictive and incremental validity was also reported to be favorable although modest. Data around acceptability, however, was equivocal; and we found no direct evidence on feasibility and

cost-effectiveness of the tool. These findings are, however, limited mostly to the United Kingdom and Australasian context.

While we found only three studies on the utility of the CPST, all based in the UK medical setting; internal reliability of the tool was reported as good. Content validity of the tool was found to be modest, and greater predictive validity was noted when the CPST is combined with the SJT. Acceptability of the tool across relevance and fairness as perceived by the applicants was reported as high. Empirical data around cost-effectiveness was lacking.

Selection centers, seemed to have favorable internal validity and predictive validity with respect to global selection score. However, concerns about acceptability, feasibility and cost-effectiveness were raised, and the data were limited to the UK and Australian GP setting.

Evidence around the psychometric robustness of other selection methods, such as speciality-specific tests, letters of recommendations, personal statements and curriculum vitae, was limited and equivocal given the paucity of studies in these methods.

### **What are the predictors of success in subsequent performance?**

#### **Range of evidence**

We discussed the predictive validity of specific selection-based tools in an earlier section. In this section, we focus on a range of predictors that have been used in locally determined selection processes. A total of 27 studies in our review reported data on various predictors of a specialist trainee's subsequent performance. The predictor variables included past academic achievement indicators such as the USMLE and other certifying and specialty-specific examinations, grades/grade-point average (GPA), rank order and Alpha Omega Alpha Medical Honors Society (AOA) status, composite selection scores, letters of recommendation and personal characteristics of applicants. Outcome variables that indicated future performance included: performance in in-training examinations, work-based assessments, faculty evaluations and end-of training examinations. Most studies were based in North America and almost all the studies employed retrospective study designs using correlation analysis and regression modeling.

#### **Predictive value of academic selection criteria**

A number of studies in our review reported the USMLE Step 1 score to be an independent predictor of resident success in terms of significantly positive correlations with both in-service and end-of-training examinations (Boyse et al. 2002; Brothers and Wetherholt 2007; Shellito et al. 2010; De Virgilio et al. 2010; Dougherty et al. 2010; Shiroma and Alarcon 2010; Alterman et al. 2011). A few studies found, however, that the USMLE Step 2 scores to be a better predictor of resident's performance than the Step 1 scores, especially towards the later years in the training (Bell et al. 2002; Thundiyil et al. 2010; Spurlock et al. 2010). Turner et al. (2006) found USMLE Step 1 scores to be statistically associated with the outcomes in the orthopedic in-training examination but not with the end-of-training certifying examination. One study (Gunderman and Jackson 2000) found no correlation though between USMLE Step 1

and II examination scores and radiology end of training examinations.

There was conflicting evidence from faculty assessments of residents' performance in core competency areas. Alterman et al. (2011) found the USMLE Step 1 to be positively correlated with the assessment scores in the core ACGME (Accreditation Council for Graduate Medical Education) competencies; and Hayden et al. (2005) reported USMLE percentile scores to correlate fairly well with the overall assessment of residents' performance. In contrast, Brothers and Wetherholt (2007) reported the USMLE to be correlated negatively with the resident clinical performance, whereas Boyse et al. (2002) and Stohl et al. (2010) found no predictable correlation of the USMLE scores with performance during residency.

#### **Predictive value of other selection criteria (grades; AOA status; research experience; gender, ethnicity and CVs)**

Evidence around the predictive value of grades and rank was equivocal. Seven articles provided some evidence of a positive association between medical school performance, cumulative grade point average and Honors or A grades, with subsequent performance during in-training evaluations and/or end-of-training examination (Boyse et al. 2002; Dirschl et al. 2002; Khongphatthanayothin et al. 2002; Hayden et al. 2005; Turner et al. 2006; Shellito et al. 2010; Selber et al. 2014). On the other hand, Brothers and Wetherholt (2007) found that while the grade point average correlated positively with the certifying examinations in general surgery, there was no association with the core competency of knowledge, and the association was negative with the performance on communication, professionalism and patient care. Similarly, Alterman et al. (2011) and Bell et al. (2002) found no association between medical school grades and the number of honors with subsequent performance. On examining the predictive validity of medical school grades, test scores, research achievements, letters of recommendations and personal statements, Stohl et al. (2010) found no significant association between these measures with subsequent performance in residency. Alterman et al. (2011) found gender and ethnicity to be non-predictive of general surgery residents' future performance.

There was similarly conflicting evidence about the Alpha Omega Alpha Medical Honors Society (AOA) status. While Dirschl et al. (2002) and Shellito et al. (2010) found AOA status to correlate positively with residency performance, Turner et al. (2006) showed that, although the AOA status correlated with the in-training examination, it did not correlate with the end-of training certifying examination. Furthermore, Boyse et al. (2002) and Alterman et al. (2011) found no correlation between AOA status and performance on in-training and end-of-training examinations. The conflicting evidence could be explained by the fact that the AOA nomination is based on academic results in combination with non-cognitive factors such as leadership and professionalism.

Other selection criteria reported as having good predictive value include prior training experience in the relevant specialty (Selber et al. 2014) and research experience (De Virgilio et al. 2010). However, there was no predictive value

of the number of research projects and publications for future performance (Dirschl et al. 2002).

### *Predictive value of composite scores*

We found some evidence that while the individual components of selection criteria may not correlate with future performance, a combined score may correlate well. Composite scores that correlated positively with future performance measures included the: Quantitative Composite Scoring Tool comprising USMLE scores, AOA status and honors grades (Turner et al. 2006), the global assessment score, comprising: interview, letters of recommendation and clinical grades (Ozuah 2002); and the total selection score (CV, referee reports, interviews) (Oldfield et al. 2013). On the contrary, Bell et al. (2002) found no predictive value of composite score based on interviews, letters of recommendation, number of honors and USMLE.

### *Predictive value of nonacademic selection criteria*

A few studies in our review examined the predictive validity of nonacademic criteria for future performance. Hayden et al. (2005) found categories of distinctive factors such as being a top-level athlete, musician and involvement in student organizations at national level to be predictors of overall success in residency. In terms of personal/behavioral characteristics of applicants as assessed during the interview, Brothers and Wetherholt (2007) found the combined score of applicants' "personal characteristics" such as attitude, motivation, integrity, interpersonal relationships and responses to specific life challenges to correlate favorably with residents' clinical performance in core competencies. On the other hand, Dawkins et al. (2005) assessed the predictive validity of psychiatric residency applicants' scores on five dimensions: empathy, academic potential, clinical potential, team-player, and an overall rating and found no association with residents' subsequent performance in terms of rotation evaluations and in-service examination scores. Similarly, Selber et al. (2014) found no predictive value of an applicant's presentation, personality, social, communication and skills as a team-player. Using a validated instrument to assess emotional intelligence (EI), Lin et al. (2013) found no correlation between the EI and various academic parameters, such as USMLE examination scores, medical school grades and AOA status. While applicant EI did correlate moderately with rank status, it did not correlate with the faculty evaluations during the selection process indicating a possible inability of the interviews to capture adequately the EI of applicants. Bohm et al. (2014) found no correlation between a validated test of moral reasoning, the Defining Issues Test 2 (DIT-2) and rank order of orthopedic surgery resident applicants.

Personality type (using Myers-Briggs-type indicator) has been suggested as an influence in selection. Quintero et al. (2009) found that there was a significant association between similarities in personality type and individual faculty interviewers' rankings of applicants. Interestingly, clinicians were prone to rate applicants of the same personality type favorably.

To summarize, most of the studies in our review exploring predictive validity of selection processes based on locally determined criteria, were based in North America.

USMLE scores were the most widely researched, and we found some evidence of USMLE Step I and II scores to be independent predictor of trainees' in-service and end-of-training examination scores. However, the evidence around USME scores' predictive value for trainees' performance in competency areas was conflicting.

Evidence of the predictive validity of other markers of academic achievement such as medical school grades, rank, research achievements, and the AOA status was equivocal although some studies found positive association of these criteria with in-training and end-of-training examinations.

In relation to the predictive value of nonacademic criteria (personality traits, communication skills, social skills etc.), it is difficult to reach a consensus due to lack of sufficient number of studies and due to conflicting findings.

## **Discussion**

### *Summary of findings*

Our findings have synthesized the evidence about the underlying frameworks of selection systems as a whole, the effectiveness of methods of selection and their predictive validity for successful performance. There was a paucity of data in illuminating our first review question on the underlying frameworks and principles of selection. There was a sense that selection frameworks have been developed in isolation to other important and related curricular concepts within medical education and training such as assessment. While there were some linkages in selection frameworks to the tenets of competency-based medical education (CBME), there was little linkage with the advances in assessment of trainees such as developments in work-based assessment (Barrett et al. 2016). Of those studies that did express a statement about underpinning concepts, most were limited to reflecting upon the need to consider both personal academic (or cognitive) and nonacademic (non-cognitive) capabilities (Patterson et al. 2008; Patterson, Tavabie, et al. 2013). We do not feel that this constitutes a framework as defined in general terms.

However, we can classify selection frameworks into two broad categories: one in which the selection criteria are locally defined, subjective, and primarily academically oriented, and the second that uses multiple methods with relatively well-defined selection criteria drawn from recognizable CBME principles (Frank et al. 2010).

The first framework that underpins selection systems, in the US for example is based on locally derived selection criteria, often viewed as subjective with substantial weighting on past academic achievement (Makdisi et al. 2011). We found that the most valued factors in the selection in this system, as perceived by local program directors, included scores in the national licensing examinations, scores from interviews, and letters of recommendation. Evidence on relative importance of other criteria such as candidates' research potential of and their nonacademic attributes was inconclusive.

The second selection framework that is gaining momentum, particularly in the United Kingdom, involves relatively well-defined selection criteria with multiple methods of selection assessing multiple skills. While the number of studies is limited and contextualized to particular settings, our review highlighted some empirical evidence toward

**Table 2.** Summary of evidence and challenges relating to various selection methods into specialty training.

Format/tools	Evidence	The challenges
MMI	Relatively high reliability for an observed assessment. Flexibility in format. Results reproducible in several settings. Favorable predictive validity.	Most MMIs have locally derived marking criteria. Data supporting validity often context-specific. Concerns raised, but not substantiated that cost-effectiveness restricts their use.
Structured/semistructured interview	Mixed evidence (moderate to high) reliability but limited generalizability to other settings.	Trained interviewees on uniform or standardized scoring systems, ideally related to the institution's training standards and frameworks required for reproducibility.
SJT	Evidence-based emerging around their use within competency-based selection systems. So far, favorable reliability and predictive validity.	Results yet to be reproduced in other settings. Concerns raised around high development costs.
CPST	Favorable reliability, although a function of the number of items sampling the underlying assessment blueprint. Favorable reliability when combined with SJT.	Little evidence they are testing problem solving, rather than acquired expected knowledge. Attractiveness lies in cost-effectiveness, and in being reproducible at reasonable levels in settings using knowledge-based assessments.
Simulations (within selection centers)	Structured marking, interviewer training and multiple tasks can assist in achieving reliability.	Multi-station, multiassessor assessments are costly to design and implement. Outside of the test development centers, there is skepticism that psychometric robustness can be achieved when using a few stations.
Letters of recommendation	Trend exists towards using structured letters of recommendation as opposed to narrative letters, but evidence on reliability and validity is limited.	Some centers claim good reliability is possible, but this has not been reproducible in other settings.
Personal statement and CVs	No firm evidence that personal statements have value in postgraduate settings. No correlation between the quality of the CV and subsequent performance was found.	CVs tend to be used in interviews in a non-standardizable way.
Predictive value of academic criteria	Trend exists towards using USMLE scores in residency selection in the United States as a measure of knowledge, and a reasonable predictor of performance on subsequent in- and end-of-training assessments. However, the evidence around USMLE scores' predictive value for trainees' performance in competency areas was conflicting.	The test not designed to be a primary determinant of the likelihood of success in residency. Uncertain consequences for applicants in moving away from holistic assessment of the skills and behaviors sought future health specialists.
Predictive value of nonacademic criteria	Little research on predictive value of aspects of personality in the context of selection into specialty training	Little justification in developing personality testing based on current frameworks.

identification and aligning of core and speciality-specific competencies for applicants and assessors to the selection criteria.

For our second review question on the utility of selection methods, we found differences in psychometrics used to interpret data specifically on the reliability of the tools. Some studies, for instance, have used raw correlations, and others have used corrected coefficients (significant). Similarly, differences in the observation-based reliability coefficients, which are reported variously as inter-rater reliability, internal structure of the measurement tool and generalizability makes comparisons difficult.

Regarding validity, many of the studies do not refer to validity frameworks or specify the particular types of validity evidence they are collecting. Differences in methods mean that it is difficult to compare studies. For example, a recent development in validity research is around consequential validity, which describes the intended and unintended effects on stakeholders of any assessment (Cook et al. 2015). Of the studies that addressing aspects of consequential validity of selection methods, one (Patterson et al. 2012) referred to the concept of political validity, and the other referred to social validity (Burgess et al. 2014) in exploring candidates' perceptions of job relevance and overall fairness of the selection process in general practice training.

Innovations in selection systems for postgraduate training in the United Kingdom, the Netherlands, Denmark, Canada and Australia are primarily referring to the CBME framework to design their selection criteria and methods.

In any field of assessment, no one method can test all the necessary attributes, thus using a combination of methods in selection broadens the range of measurable attributes (Patterson and Ferguson 2012). These include multiple mini-interviews, situational judgment tests, clinical problem-solving tests and their combinations. Table 2 summarizes evidence on the utility of these tools including design and implementation challenges.

In regards to interviews, we found MMIs to have favorable inter-rater reliability, acceptability and predictive validity of end-of-training scores; however, there was conflicting evidence about what MMIs were testing, that is, issues with their construct validity. Feasibility appears problematic in terms of resource implications; however, on comparing the feasibility of MMIs with traditional interviews, there was a recognition that MMIs need more planning in terms of physical resources and personnel involved, but this may be an issue during the initial set-up rather than ongoing maintenance. MMIs have been considered an acceptable way to assess characteristics such as professionalism in a high-stakes decision (Hofmeister et al. 2008). For both structured interviews as well as MMIs, it is important that sufficient information is given to the applicants in advance (Isaksen et al. 2013) and that interviewers have had appropriate training, although this strategy by itself may not be able to account for differences in interviewer reliability (Roberts et al. 2014).

Apart from MMIs and semistructured interviews, the literature on personnel selection in human resources reports several other labels such as "situational", "behavioral",



“conventional structured” and “structured situational” interviews (Macan 2009). However, there seems to be a paucity of evidence in medical education around the use of other types of structured interviews in selection.

There is a good level of consensus from a range of evidence in our review to support the use of the situational judgment test (SJT) as an element of postgraduate selection systems. It has been found to have good internal consistency, with the caveat of test specification and construction being demonstrated mostly in the original development setting. We found the SJT to have favorable criterion validity, and it was a modest predictor of end-of-training scores. However, there has been little consideration of its construct validity, that is, what it is testing, a problem shared by the MMI. In the United Kingdom, there appears to be an overall incremental improvement in evidencing validity from the initial pilot-testing to operation of the SJT as a standard test format in postgraduate selection systems. Generally, SJTs are complex to develop and there is a wide range of options available in relation to item formats, instructions and scoring. However, with the quality improvements in the testing specifications and overall experience of the SJT, its application in other international postgraduate training settings is expected to improve. Given the increasing pressure on external accountability and cost-efficiency in postgraduate training internationally, it may be desirable to use computer-based technologies. More valid versions of the SJT relying on such technology have not been extensively trialed. We could not find any empirical published data on costings of the SJT or the MMI, although the latter has been costed in undergraduate settings. In an effort to reduce costs associated with mounting an MMI at an international site for international applicants, (Tiller et al. 2013) introduced an Internet-based iMMI that utilized Skype. Favorable findings were reported for the iMMI in terms of reliability, validity, acceptability and savings of resources. In Germany, costs of the undergraduate MMI were \$485 per applicant (Hissbach et al. 2014). Wakeford suggested that multiple choice tests such as the clinical problem-solving test would cost from \$125–250 per applicant but selection centers, running simulations might cost \$1250 per applicant (Wakeford 2014).

Regarding the Clinical Problem-Solving Test (CPST), where test specifications fit in with an overarching competency-based framework, as in the UK general practice setting, predictive validity of the CPST and SJT combined seems encouraging, but its reproducibility in other less integrated selection systems needs more research.

Selection centers appear attractive in combining tests to assess a greater range of entry-level attributes. However, the literature suggests that this concept needs further theoretical development as the label can apply to other simulation exercises or a combination of results derived from a programmatic selection process that is not necessarily conducted in the same physical location.

The evidence around the utility of other selection methods such as letters of recommendation, personal statements, and CVs were inadequate to make any judgmental claims. In relation to letters of recommendation, while the “standardized” format was found to be more feasible and acceptable, caution should be displayed in relation to the potential for the “standardized” letters to lead to inflated scores for applicants, particularly when the letter-writer is

less experienced with reference-writing or has had less experience with the applicant.

Similarly, we found a lack of sufficient papers on assessment of specialty-specific skills. We found only two papers on technical skills testing such as surgery-specific skills. Similarly, only two papers examined nonacademic skills (emotional intelligence and moral reasoning) among applicants. The theoretical fit and specifications of such tests need to be linked in with an overarching selection framework, particularly when attributes that constitute the affective domain such as empathy and perspective-taking, integrity, reliability, diligence, trustworthiness, commitment, respect and interpersonal skills have been acknowledged as important skills for the competent practitioner (Bernstein et al. 2003; Patterson, Tavabie, et al. 2013).

In terms of our third review question regarding, the predictors of success in subsequent performance, the bulk of evidence is around the predictive value of factors that reflect past academic achievement. Since most of the studies based in the United States, the USMLE scores have been widely researched for their predictive value in subsequent performance in in-training as well as end-of-training examinations and faculty assessments of residents. While the USMLE scores were found to correlate well with in-service and end-of-training examination scores, the evidence was inconclusive in relation to their predictive value in faculty’s assessment of residents in core competency areas. These findings bring to the fore concerns that while past academic scores are good indicators of future academic/cognitive scores, they do not indicate success in a trainee’s overall performance that goes beyond cognitive capabilities (Stohl et al. 2010). The increasing use of the USMLE Step 1 component to screen applicants for residency has increased despite the test not being designed as a primary determinant of the likelihood of success in residency. This is likely to lead to unintended consequences for students and universities who seek to alter curricula (Prober et al., 2016). However, it is unclear who will bear the cost of developing a holistic assessment of the skills, attributes, and behaviors sought in future health care providers.

### Implications

Owing to the multidimensional and complex role of a specialist, one of the major challenges of researching selection systems at postgraduate level is to develop a consensus on the expected generic and discipline specific competencies of a specialist. While in some locations globally, postgraduate medical education is undergoing the paradigm shift towards competency-based approaches to design and implementation of training curricula (Frank et al. 2010), discordance still exists in several other selection systems when linking this approach in developing selection systems.

The majority of studies in our review focused on psychometric properties of specific selection methods with the highest regard being given to predictive validity and the most desired endpoint in terms of subsequent within or end of training performance. Those methods which have the strongest evidence included the MMI and the SJT. Findings of such studies have led to important advances in selection-focused assessment and have provided good evidence about the strengths and weaknesses of the various

approaches as well as understandings of their relationships. However given the important cost considerations, independent study into the cost effectiveness of the MMI and the SJT formats is required.

It seems that researchers have been diligent in making the best use of secondary analysis of data in reporting simple correlations between variables of interest or using regression to see which selection methods predict future performance. Nevertheless, concerns about the reductionist underpinnings of such analyses have been raised in the wider literature especially that they do not capture the authenticity of real life (Prideaux et al. 2011). There are a number of common methodological issues that are rarely acknowledged in the predictive validity research. One of the issues is common method variance: 'tests predicting tests' between trainee selection scores and in-training assessments, as the applicants have all been selected to have the same high-end characteristics. Secondly, there is the issue of disattenuation, which takes into account measurement errors of some of the variables of interest. Furthermore, the 'latency problem', which describes the interval between point measurements, for example between selection and end-of-training assessments, may confound the stated statistical associations as reported correlations may be low or modest. If the higher number is the best from a measurement perspective, one can understand why it is tempting to use a national licensing examination such as the USMLE as the single best predictor into residency (Prober et al. 2016). This is despite leading medical educators pointing out that it is unsuitable for such a use. Rather we recommend there should be a focus on multi-method programmatic approaches in collecting, analyzing, interpreting and reporting data from a range of instruments that are fit for purpose. These rules could be reasonably reported as a global consensus so that future research about the differing selection systems is reported in a way that can be comparable.

Given its nonlinear, and dynamic nature, specialist training environments can be deemed as complex and complicated (Glouberman and Zimmerman 2002). Other than contextualized competency-based training approaches, there was no evidence in the literature of addressing issues around complexity of specialist training environments including change management issues when introducing a competency-based model of selection.

Another locally based approach to finding new predictors of success is the use of big data to inform selection decisions. The use of professionally and nonprofessionally oriented social networking web sites such as LinkedIn and Facebook have become widespread in employee recruitment and selection especially in business sectors (Nikolaou 2014). Some researchers (Go et al. 2012; Shin et al. 2013) have explored the potential of harvesting data from social media platforms to capture nonacademic data while screening or shortlisting applicants.

## Conclusions

The quality of high-stakes selection processes can be much improved if the system is based on the principles of good assessment in the context of complex specialty training environments within modern healthcare. Internationally,

*laissez-faire* approaches to locally defined selection systems as prevalent in the United States are giving way to the systematic introduction and evidencing of competency-based training approaches to selection in for example, the United Kingdom, the Netherlands and Australia. The evidence in specialty selection confirms the important advances in selection-focused assessment, with some good evidence about the strengths and weaknesses of the various approaches as well as understandings of their relationships. While much has been gained in the utility of a range of selection formats, there are many assumptions about the underlying theoretical and conceptual frameworks that are yet to be investigated.

Moving to a theory-informed research process including analysis of systemic changes brought about by the introduction of new selection systems will ensure that selection research will move beyond focus on test formats to one which explores critical questions around the consequences on applicants, training programs, and the wider community which future specialists will serve.

## Strengths and limitations of the review

The strength of this study is that we identified and synthesized the evidence that underpins the design, implementation and evaluation of selection into specialty training. Regarding gaps, the findings of the review should be interpreted against limitations on the quality as well as quantity of evidence which constrained our analysis. The majority of studies were either contextualized in North America (predictors of success) or United Kingdom (MMI, SJTs and selection centers). We were unable to include gray literature due to a lack of its availability in the public domain. We also encountered difficulty in developing comprehensive and effective search syntax due to enormous variations in search terms. The exclusion of public health and occupational and environmental health is also acknowledged. Inability to conduct a meta-analysis of quantitative data due to differences in context and reporting outcomes is also one of the limitations of this review.

## Recommendations for further research

Given the variation in specialty training across the globe and substantive gaps in the literature, selection frameworks needs significant reframing. We suggest a range of priorities that might guide the postgraduate selection agenda:

### Developing holistic selection frameworks

Competency-based frameworks are an advance over *laissez-faire* or locally defined systems as selection in such frameworks is viewed as one high-stakes assessment in the broader schema of training and lifelong learning. Selection frameworks can be strengthened by judicious use of job analysis techniques and ethnographic methods involving stakeholders to provide insights into what constitutes best practice in specialty selection are most likely formed, contested and legitimized or reformed (Stacey 1996).

## Addressing the change management agenda for implementing selection approaches

A clear gap in the evidence-base is implementation and evaluation of selection approaches from the systemic perspectives of change management principles, drawn from the broader literature including sociology, social psychology and organization management literatures. Similar to any organizational innovations, changes in selection is contingent on many other criteria beyond the psychometric qualities of specific selection methods. Success of organizational innovations is reliant on how impacted individuals and organizations will “talk the innovation up or down,” their receptiveness to new thinking about what constitutes innovation and best practice in a field (Clegg and Matos 2017), as well as their perception of the change in terms of organizational framing (Bolman and Deal 1991). A deeper and theoretical critical analysis of the circumstances concerning the impacts of changes to practices, processes and outcomes would be a valuable contribution to the literature on selection into specialist medical training.

## Maintaining diversity of the workforce

Further research on specialty selection could consider for whom the innovations in selection are designed, whose interests they serve and who they marginalize? For example, there has been research into the impact of national selection systems on uptake of rural training (Sureshkumar et al. 2017). This raises the vital question of ensuring equity in selection so that the cultural background of doctors is representative of the community they serve (Betancourt et al. 2003) as well as contribute to broader social justice agendas through widening participation in the medical workforce (Sullivan 2004).

## Broadening the scope of research methods

Reframing the selection research agenda beyond psychometric models will allow us to build research around important research questions, rather than on traditional methodological and conventional preferences. There are several promising approaches that can guide an enrich research agendas such as such as theoretical developments in multi-method programmatic approaches in collecting, analyzing, and reporting data from a range of observations that are fit for purpose (Schuwirth and Van der Vleuten 2011). Given that selection is a critical moment of assessment in transitioning from one level of training to the next, it is imperative to form synergies between frameworks and methods connecting selection, work-based assessment, and end of training assessments.

## Acknowledgements

The review team would sincerely like to thank the information scientist, Mr Lars Eriksson at the School of Medicine Library, University of Queensland, Australia for support in framing and executing the search strategy. The team would also like to thank the Royal Australasian College of Physicians for providing support and protected time for undertaking this review.

## Disclosure statement

CR: consultancy for the RACP and the AGPT/DOH on matters of selection into postgraduate training.

## Notes on contributors

**Chris Roberts**, MBChB, FRACGP, MMedSci, PhD, is an Associate Professor, Medical Education at the University of Sydney, New South Wales, Australia.

**Priya Khanna**, MSc, MEd, PhD, is a researcher at the Royal Australasian College of Physicians, New South Wales, Australia.

**Louise Rigby**, is a PhD candidate at University of Sydney is a manager at the Health Education and Training Institute, New South Wales, Australia.

**Emma Bartle**, PhD, is Teaching and Learning Chair, School of Dentistry at the University of Queensland, Australia.

**Anthony Llewellyn**, BMedSci, MBBS, FRANZCP, MHA, GAICD, is a senior staff specialist in psychiatry training, Hunter New England Local Health District, New South Wales, a senior lecturer, University of Newcastle, and a specialist lead in Rural, Health Education and Training Institute, New South Wales, Australia.

**Julie Gustavs**, PhD, is a manager at the Royal Australasian College of Physicians, New South Wales, Australia.

**Libby Newton**, BSc, is a researcher at the Royal Australasian College of Physicians, New South Wales, Australia.

**James P. Newcombe**, BMedSci (Hons), MPH (Hons), MBBS, GAICD, FRACP, FRCPA, is an infectious diseases physician and clinical microbiologist at the Royal North Shore Hospital, New South Wales, Australia.

**Mark Davies**, MBBS, FRACP, is a staff specialist in neonatology at the Royal Brisbane and Women's Hospital and an associate professor of neonatology at the University of Queensland, Australia.

**Jill Thistlethwaite**, BSc, MM, MS, PhD, MMed, FRCGP, FRACGP, is an adjunct professor at University Technology Sydney, honorary professor in the School of Education at the University of Queensland, and a medical advisor to the NPS MedicineWise in Australia.

**James Lynam**, BSc (Hons), MBBS, MRCP, FRACP, is a practicing medical oncologist at the Calvary Mater Newcastle, the Network Director of Physician Training for the Hunter New England Network and a conjoint lecturer at the University of Newcastle, New South Wales, Australia.

## References

- Adusumilli S, Cohan RH, Marshall KW, Fitzgerald JT, Oh MS, Gross BH, Ellis JH. 2000. How well does applicant rank order predict subsequent performance during radiology residency? *Acad Radiol.* 7:635–640.
- Ahmed A, Abid MA, Bhatti NI. 2017. Balancing standardized testing with personalized training in surgery. *Adv Med Educ Pract.* 8:25.
- Ahmed A, Qayed KI, Abdulrahman M, Tavares W, Rosenfeld J. 2014. The multiple mini-interview for selecting medical residents: first experience in the Middle East region. *Med Teach.* 36:703–709.
- Ahmed H, Rhydderch M, Matthews P. 2012a. Can knowledge tests and situational judgement tests predict selection centre performance? *Med Educ.* 46:777–784.
- Ahmed H, Rhydderch M, Matthews P. 2012b. Do general practice selection scores predict success at MRCGP? An exploratory study. *Educ Primary Care.* 23:95–100.
- Al Khalili K, Chalouhi N, Tjoumakaris S, Gonzalez LF, Starke RM, Rosenwasser R, Jabbour P. 2014. Programs selection criteria for neurological surgery applicants in the United States: a national survey for neurological surgery program directors. *World Neurosurg.* 81:473–477.
- Alterman DM, Jones TM, Heidel RE, Daley BJ, Goldman MH. 2011. The predictive value of general surgery application data for future resident performance. *J Surg Educ.* 68:513–518.
- Andrades M, Bhanji S, Kausar S, Majeed F, Pinjani S. 2014. Multiple mini-interviews (MMI) and semistructured interviews for the selection of family medicine residents: a comparative analysis. *Int Sch Res Notices.* 2014:1.
- Baldwin K, Weidner Z, Ahn J, Mehta S. 2009. Are away rotations critical for a successful match in orthopaedic surgery?. *Clin Orthop Relat Res.* 467:3340–3345.



- Bandiera G, Regehr G. 2004. Reliability of a structured interview scoring instrument for a Canadian postgraduate emergency medicine training program. *Acad Emerg Med.* 11:27–32.
- Barrett A, Galvin R, Steinert Y, Scherpbier A, O'Shaughnessy A, Horgan M, Horsley T. 2016. A BEME (Best Evidence in Medical Education) review of the use of workplace-based assessment in identifying and remediating underperformance among postgraduate medical trainees: BEME Guide No. 43. *Med Teach.* 38:1188–1198.
- Bell JG, Kanellitsas I, Shaffer L. 2002. Selection of obstetrics and gynecology residents on the basis of medical school performance. *Am J Obstet Gynecol.* 186:1091–1094.
- Bernstein AD, Jazrawi LM, Elbesheshy B, Valle CJD, Zuckerman JD. 2003. An analysis of orthopaedic residency selection criteria. *Bull Hosp Jt Dis.* 61:49–57.
- Beskind DL, Hiller KM, Stolz U, Bradshaw H, Berkman M, Stoneking LR, Fiorello A, Min A, Viscusi C, Grall KJ. 2014. Does the experience of the writer affect the evaluative components on the standardized letter of recommendation in emergency medicine? *J Emerg Med.* 46:544–550.
- Betancourt JR, Green AR, Carrillo JE, Ananeh-Firempong O II, 2003. Defining cultural competence: a practical framework for addressing racial/ethnic disparities in health and health care. *Public Health Rep.* 118:293.
- Bohm KC, Heest TV, Gioe TJ, Agel J, Johnson TC, Heest AV. 2014. Assessment of moral reasoning skills in the orthopaedic surgery resident applicant. *J Bone Joint Surg Am.* 96:e151.
- Bolman LG, Deal TE. 1991. Leadership and management effectiveness: a multi-frame, multi-sector analysis. *Hum Resour Manage.* 30:509–534.
- Boyse TD, Patterson SK, Cohan RH, Korobkin M, Fitzgerald JT, Oh MS, Gross BH, Quint DJ. 2002. Does medical school performance predict radiology resident performance? *Acad Radiol.* 9:437–445.
- Brothers TE, Wetherholt S. 2007. Importance of the faculty interview during the resident application process. *J Surg Educ.* 64:378–385.
- 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.
- Cameron AJ, Mackeigan LD, Mitsakakis N, Pugsley JA. 2017. Multiple mini-interview predictive validity for performance on a pharmacy licensing examination. *Med Educ.* 51:379–389.
- Campagna-Vaillancourt M, Manoukian J, Razack S, Nguyen LH. 2014. Acceptability and reliability of multiple mini interviews for admission to otolaryngology residency. *Laryngoscope.* 124:91–96.
- Carroll SM, Kennedy A, Traynor O, Gallagher AG. 2009. Objective assessment of surgical performance and its impact on a national selection programme of candidates for higher surgical training in plastic surgery. *J Plast Reconstr Aesthet Surg.* 62:1543–1549.
- Chew FS, Ochoa ER, Relyea-Chew A. 2005. Spreadsheet application for radiology resident match rank list 1. *Acad Radiol.* 12:379–384.
- Clegg SR, Matos J. 2017. *Sustainability and organizational change management.* Routledge.
- Colquitt JA, Conlon DE, Wesson MJ, Porter CO, Ng KY. 2001. Justice at the millennium: a meta-analytic review of 25 years of organizational justice research. *J Appl Psychol.* 86:425.
- Cook DA, Beckman TJ. 2006. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med.* 119:166e7–166e16.
- Cook DA, Brydges R, Ginsburg S, Hatala R. 2015. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ.* 49:560–575.
- Crane JT, Ferraro CM. 2000. Selection criteria for emergency medicine residency applicants. *Acad Emerg Med.* 7:54–60.
- Crossingham G, Gale T, Roberts M, Carr A, Langton J, Anderson I. 2011. Content validity of a clinical problem solving test for use in recruitment to the acute specialties. *Clin Med.* 11:22–25.
- Davison I, Burke S, Bedward J, Kelly S. 2006. Do selection scores for general practice registrars correlate with end of training assessments? *Educ Prim Care.* 17:473.
- Dawkins K, Ekstrom RD, Maltbie A, Golden RN. 2005. The relationship between psychiatry residency applicant evaluations and subsequent residency performance. *Acad Psychiatr.* 29:69–75.
- De Virgilio C, Yaghoubian A, Kaji A, Collins JC, Deveney K, Dolich M, Easter D, Hines OJ, Katz S, Liu T. 2010. Predicting performance on the American Board of Surgery qualifying and certifying examinations: a multi-institutional study. *Arch Surg.* 145:852–856.
- Dirschl DR, Dahners LE, Adams GL, Crouch JH, Wilson FC. 2002. Correlating selection criteria with subsequent performance as residents. *Clin Orthop Relat Res.* 399:265–274.
- Dore KL, Kreuger S, Ladhani M, Rolfson D, Kurtz D, Kulasegaram K. 2010. The reliability and acceptability of the multiple mini-interview as a selection instrument for postgraduate admissions. *Acad Med.* 85:S60–S63.
- Dougherty PJ, Walter N, Schilling P, Najibi S, Herkowitz H. 2010. Do scores of the USMLE Step 1 and OITE correlate with the ABOS Part I certifying examination?: a multicenter study. *Clin Orthop Relat Res.* 468:2797–2802.
- Egol KA, Collins J, Zuckerman JD. 2011. Success in orthopaedic training: resident selection and predictors of quality performance. *J Am Acad Orthop Surg.* 19:72–80.
- Eva KW, Macala C. 2014. Multiple mini-interview test characteristics: 'tis better to ask candidates to recall than to imagine. *Med Educ.* 48:604–613.
- Eva KW, Reiter HI, Trinh K, Wasi P, Rosenfeld J, Norman GR. 2009. Predictive validity of the multiple mini-interview for selecting medical trainees. *Med Educ.* 43:767–775.
- Eva KW, Rosenfeld J, Reiter HI, Norman GR. 2004. An admissions OSCE: the multiple mini-interview. *Med Educ.* 38:314–326.
- Farkas DT, Nagpal K, Curras E, Shah AK, Cosgrove JM. 2012. The use of a surgery-specific written examination in the selection process of surgical residents. *J Surg Educ.* 69:807–812.
- Ferguson E, James D, Madeley L. 2002. Factors associated with success in medical school: systematic review of the literature. *BMJ.* 324:952–957.
- Frank JR, Danoff D. 2007. The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. *Med Teach.* 29:642–647.
- Frank JR, Snell LS, Cate OT, Holmboe ES, Carraccio C, Swing SR, Harris P, Glasgow NJ, Campbell C, Dath D, et al. 2010. Competency-based medical education: theory to practice. *Med Teach.* 32:638–645.
- Fraser JD, Aguayo P, Peter SS, Ostlie DJ, Holcomb GW, III, Andrews WA, Murphy JP, Sharp RJ, Snyder CL. 2011. Analysis of the pediatric surgery match: factors predicting outcome. *Pediatr Surg Int.* 27:1239–1244.
- Gale T, Roberts M, Sice P, Langton J, Patterson F, Carr A, Anderson I, Lam W, Davies P. 2010. Predictive validity of a selection centre testing non-technical skills for recruitment to training in anaesthesia. *Br J Anaesth.* 105:603–609.
- Gallagher AG, Neary P, Gillen P, Lane B, Whelan A, Tanner WA, Traynor O. 2008. Novel method for assessment and selection of trainees for higher surgical training in general surgery. *ANZ J Surg.* 78:282–290.
- Glouberman S, Zimmerman B. 2002. Complicated and complex systems: what would successful reform of Medicare look like? *Romanow Papers.* 2:21–53.
- Go PH, Klaassen Z, Chamberlain RS. 2012. Attitudes and practices of surgery residency program directors toward the use of social networking profiles to select residency candidates: a nationwide survey analysis. *J Surg Educ.* 69:292–300.
- Gunderman RB, Jackson VP. 2000. Are NBME examination scores useful in selecting radiology residency candidates? *Acad Radiol.* 7:603–606.
- Haig A, Dozier M. 2003. BEME Guide no 3: systematic searching for evidence in medical education—Part 1: Sources of information. *Med Teach.* 25:352–363.
- Hamdy H, Prasad K, Anderson MB, Scherpbier A, Williams R, Zwierstra R, Cuddihy H. 2006. BEME systematic review: predictive values of measurements obtained in medical schools and future performance in medical practice. *Med Teach.* 28:103–116.
- Hayden SR, Hayden M, Gamst A. 2005. What characteristics of applicants to emergency medicine residency programs predict future success as an emergency medicine resident? *Acad Emerg Med.* 12:206–210.
- Hissbach JC, Sehner S, Harendza S, Hampe W. 2014. Cutting costs of multiple mini-interviews—changes in reliability and efficiency of the Hamburg medical school admission test between two applications. *BMC Med Educ.* 14:54.
- Hofmeister M, Lockyer J, Crutcher R. 2008. The acceptability of the multiple mini interview for resident selection. *Fam Med.* 40:734–740.



- Hofmeister M, Lockyer J, Crutcher R. 2009. The multiple mini-interview for selection of international medical graduates into family medicine residency education. *Med Educ.* 43:573–579.
- Hopson LR, Burkhardt JC, Stansfield RB, Vohra T, Turner-Lawrence D, Losman ED. 2014. The multiple mini-interview for emergency medicine resident selection. *J Emerg Med.* 46:537–543.
- Isaksen JH, Hertel NT, Kjaer NK. 2013. Semi-structured interview is a reliable and feasible tool for selection of doctors for general practice specialist training. *Danish Med J.* 60:A4692–A4692.
- Janis JE, Hatfield DA. 2008. Resident selection protocols in plastic surgery: a national survey of plastic surgery program directors. *Plas Reconstr Surg.* 122:1929–1939.
- Katsufrakakis PJ, Uhler TA, Jones LD. 2016. The residency application process: Pursuing improved outcomes through better understanding of the issues. *Acad Med.* 91:1483–1487.
- Khongphatthanayothin A, Chongsrisawat V, Wanankul S, Sanpavat S. 2002. Resident recruitment: what are good predictors for performance during pediatric residency training? *J Med Assoc Thai.* 85Suppl 1:S302–S311.
- 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.* 46:399–408.
- Krauss E, Bezuhly M, Williams J. 2015. Selecting the best and brightest: a comparison of residency match processes in the United States and Canada. *Plast Surg (Oakv).* 23:225.
- Lansford CD, Fisher SR, Ossoff RH, Chole RA. 2004. Otolaryngology–head and neck surgery residency match: applicant survey. *Arch Otolaryngol Head Neck Surg.* 130:1017–1023.
- Lillis S. 2010. Do scores in the selection process for vocational general practice training predict scores in vocational examinations? *Prim Health Care.* 1:114–118.
- Lin DT, Kannappan A, Lau JN. 2013. The assessment of emotional intelligence among candidates interviewing for general surgery residency. *J Surg Educ.* 70:514–521.
- Love JN, Delorio NM, Ronan-Bentle S, Howell JM, Doty CI, Lane DR, Hegarty C, Burton J. 2013. Characterization of the Council of Emergency Medicine Residency Directors' standardized letter of recommendation in 2011–2012. *Acad Emerg Med.* 20:926–932.
- Macan T. 2009. The employment interview: a review of current studies and directions for future research. *Hum Resour Manage Rev.* 19:203–218.
- Makdasi 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.* 68:67–72.
- Maverakis E, Li CS, Alikhan A, Lin TC, Idriss N, Armstrong AW. 2012. The effect of academic “misrepresentation” on residency match outcomes. *Dermatol Online J.* 18.
- Max BA, Gelfand B, Brooks MR, Beckerly R, Segal S. 2010. Have personal statements become impersonal? An evaluation of personal statements in anesthesiology residency applications. *J Clin Anesth.* 22:346–351.
- McGaghie WC, Cohen ER, Wayne DB. 2011. Are United States medical licensing exam step 1 and 2 scores valid measures for postgraduate medical residency selection decisions? *Acad Med.* 86:48–52.
- Melendez MM, Xu X, Sexton TR, Shapiro MJ, Mohan EP. 2008. The importance of basic science and clinical research as a selection criterion for general surgery residency programs. *J Surg Educ.* 65:151–154.
- Mitchison H. 2009. Assessment centres for core medical training: how do the assessors feel this compares with the traditional interview? *Clin Med.* 9:147–150.
- Moore EJ, Price DL, Abel KMV, Carlson ML. 2015. Still under the microscope: Can a surgical aptitude test predict otolaryngology resident performance? *Laryngoscope.* 125:E57–E61.
- Nguyen AT, Janis JE. 2012. Resident selection protocols in plastic surgery: a national survey of plastic surgery independent program directors. *Plast Reconstr Surg.* 130:459–469.
- Nikolaou I. 2014. Social networking web sites in job search and employee recruitment. *Int J Select Assess.* 22:179–189.
- 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.* 18:310–315.
- Oldfield Z, Beasley SW, Smith J, Anthony A, Watt A. 2013. Correlation of selection scores with subsequent assessment scores during surgical training. *ANZ J Surg.* 83:412–416.
- Ozuah PO. 2002. Predicting residents' performance: a prospective study. *BMC Med Educ.* 2:1.
- Pashayan N, Gray S, Duff C, Parkes J, Williams D, Patterson F, Koczwara A, Fisher G, Mason B. 2016. 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.* 38:e194.
- Patterson F, Baron H, Carr V, Plint S, Lane P. 2009. Evaluation of three short-listing methodologies for selection into postgraduate training in general practice. *Med Educ.* 43:50–57.
- Patterson F, Carr V, Zibarras L, Burr B, Berkin L, Plint S, Irish B, Gregory S. 2009. New machine-marked tests for selection into core medical training: evidence from two validation studies. *Clin Med.* 9:417–420.
- Patterson F, Ferguson E. 2010. Selection for medical education and training. Wiley-Blackwell.
- Patterson F, Ferguson E. 2012. Testing non-cognitive attributes in selection centres: how to avoid being reliably wrong. *Med Educ.* 46:240.
- Patterson F, Ferguson E, Thomas S. 2008. Using job analysis to identify core and specific competencies: implications for selection and recruitment. *Med Educ.* 42:1195–1204.
- Patterson F, Knight A, Dowell J, Nicholas S, Cousans F, Cleland J. 2016. How effective are selection methods in medical education? A systematic review. *Med Educ.* 50:36–60.
- 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:e734–e741.
- 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. *Int J Select Assess.* 20:486–496.
- 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.* 16:1–8.
- 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. *Br J Gen Pract.* 63:e331–e338.
- 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.* 38:3–17.
- Patterson F, Zibarras L, Carr V, Irish B, Gregory S. 2011. Evaluating candidate reactions to selection practices using organisational justice theory. *Med Educ.* 45:289–297.
- Patterson F, Zibarras L, Kerrin M, Lopes S, Price R. 2014. Development of competency models for assessors and simulators in high-stakes selection processes. *Med Teach.* 36:1082–1085.
- Pau A, Jeevaratnam K, Chen YS, Fall AA, Khoo C, Nadarajah VD. 2013. The Multiple mini-interview (MMI) for student selection in health professions training—a systematic review. *Med Teach.* 35:1027.
- Perkins JN, Liang C, McFann K, Abaza MM, Streubel SO, Prager JD. 2013. Standardized letter of recommendation for otolaryngology residency selection. *Laryngoscope.* 123:123–133.
- Prager JD, Perkins JN, McFann K, Myer CM, Pensak ML, Chan KH. 2012. Standardized letter of recommendation for pediatric fellowship selection. *Laryngoscope.* 122:415–424.
- Prideaux D, Roberts C, Eva K, Centeno A, McCrorie P, McManus C, Patterson F, Powis D, Tekian A, 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.* 33:215–223.
- Prober CG, Kolars JC, First LR, Melnick DE. 2016. A plea to reassess the role of United States Medical Licensing Examination Step 1 scores in residency selection. *Acad Med.* 91:12–15.
- Quintero AJ, Segal LS, King TS, Black KP. 2009. The personal interview: assessing the potential for personality similarity to bias the selection of orthopaedic residents. *Acad Med.* 84:1364–1372.
- 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.* 14:1.
- Roberts C, Togno JM. 2011. Selection into specialist training programs: an approach from general practice. *Med J Aust.* 194:93–95.

- Roberts C, Walton M, Rothnie I, Crossley J, Lyon P, Kumar K, Tiller D. 2008. Factors affecting the utility of the multiple mini-interview in selecting candidates for graduate-entry medical school. *Med Educ.* 42:396–404.
- Roberts M, Gale T, Sice P, Anderson I. 2013. The relative reliability of actively participating and passively observing raters in a simulation-based assessment for selection to specialty training in anaesthesia. *Anaesthesia.* 68:591–599.
- Robinson SW, Roberts N, Dzara K. 2013. Residency-coordinator perceptions of psychiatry residency candidates: a pilot study. *Acad Psychiatry.* 37:265–267.
- Rogers CR, Gutowski KA, Munoz DEL, Rio A, Larson DL, Edwards M, Hansen JE, Lawrence WT, Stevenson TR, Bentz ML. 2009. Integrated plastic surgery residency applicant survey: characteristics of successful applicants and feedback about the interview process. *Plast Reconstr Surg.* 123:1607–1617.
- Sbicca JA, Gorell ES, Kanzler MH, Lane AT. 2010. The integrity of the dermatology National Resident Matching Program: results of a national study. *J Am Acad Dermatol.* 63:594–601.
- Schaverien MV. 2016. Selection for surgical training: an evidence-based review. *J Surg Educ.* 73:721–729.
- Scherl SA, Lively N, Simon MA. 2001. Initial review of electronic residency application service charts by orthopaedic residency faculty members. *J Bone Joint Surg Am.* 83:65–65.
- Schuwirth LW, Van Der Vleuten CP. 2011. Programmatic assessment: from assessment of learning to assessment for learning. *Med Teach.* 33:478–485.
- Selber JC, Tong W, Koshy J, Ibrahim A, Liu J, Butler C. 2014. Correlation between trainee candidate selection criteria and subsequent performance. *J Am Coll Surg.* 219:951–957.
- Shellito JL, Osland JS, Helmer SD, Chang FC. 2010. American Board of Surgery examinations: can we identify surgery residency applicants and residents who will pass the examinations on the first attempt? *Am J Surg.* 199:216–222.
- Shin NC, Ramoska EA, Garg M, Rowh A, Nyce D, Deroos F, Carter M, Hall RV, Lopez BL, Directors DVERP. 2013. Google Internet searches on residency applicants do not facilitate the ranking process. *J Emerg Med.* 44:995–998.
- Shiroma PR, Alarcon RD. 2010. Selection factors among international medical graduates and psychiatric residency performance. *Acad Psychiatr.* 34:128–131.
- Sklar DP. 2016. Who's the fairest of them all? Meeting the challenges of medical student and resident selection. *Acad Med.* 91:1465–1467.
- Soares WE, III, Sohoni A, Hern HG, Wills CP, Alter HJ, Simon BC. 2015. Comparison of the multiple mini-interview with the traditional interview for US emergency medicine residency applicants: a single-institution experience. *Acad Med.* 90:76–81.
- Spurlock DR, Holden C, Hartranft T. 2010. Using United States Medical Licensing Examination®(USMLE) examination results to predict later in-training examination performance among general surgery residents. *J Surg Educ.* 67:452–456.
- Stacey RD. 1996. Complexity and creativity in organizations. San Francisco (CA): Berrett-Koehler Publishers.
- Stohl HE, Hueppchen NA, Bienstock JL. 2010. Can medical school performance predict residency performance? Resident selection and predictors of successful performance in obstetrics and gynecology. *J Grad Med Educ.* 2:322–326.
- Stratman EJ, Ness RM. 2011. Factors associated with successful matching to dermatology residency programs by reapplicants and other applicants who previously graduated from Medical School. *Arch Dermatol.* 147:196–202.
- Sullivan LW. 2004. Missing persons: minorities in the health professions, a report of the Sullivan Commission on Diversity in the Healthcare Workforce.
- Sureshkumar P, Roberts C, Clark T, Jones M, Hale R, Grant M. 2017. Factors related to doctors' choice of rural pathway in general practice specialty training. *Australian J Rural Health.* 25:148–154.
- Thomas H, Taylor CA, Davison I, Field S, Gee H, Grant J, Malins A, Pendleton L, Spencer E. 2012. National Evaluation of Specialty Selection: final report.
- Thundiyil JG, Modica RF, Silvestri S, Papa L. 2010. Do United States Medical Licensing Examination (USMLE) scores predict in-training test performance for emergency medicine residents? *J Emerg Med.* 38:65–69.
- Tiller D, O'Mara D, Rothnie I, Dunn S, Lee L, Roberts C. 2013. Internet-based multiple mini-interviews for candidate selection for graduate entry programmes. *Med Educ.* 47:801–810.
- Tolan AM, Kaji AH, Quach C, Hines OJ, De Virgilio C. 2010. The electronic residency application service application can predict Accreditation Council for Graduate Medical Education competency-based surgical resident performance. *J Surg Educ.* 67:444–448.
- Turner NS, Shaughnessy WJ, Berg EJ, Larson DR, Hanssen AD. 2006. A quantitative composite scoring tool for orthopaedic residency screening and selection. *Clin Orthopaed Related Res.* 449:50–55.
- Van Der Vleuten CP. 1996. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ.* 1:41–67.
- Vermeulen MI, Tromp F, Zuithoff NP, Pieters RH, Damoiseaux RA, Kuyvenhoven MM. 2014. A competency based selection procedure for Dutch postgraduate GP training: a pilot study on validity and reliability. *Eur J Gen Pract.* 20:307–313.
- Wakeford R. 2014. Predictive validity of selection for entry into post-graduate training in general practice. *Br J Gen Pract.* 64:71–71.
- Yoshimura H, Kitazono H, Fujitani S, Machi J, Saiki T, Suzuki Y, Ponnampereuma G. 2015. Past-behavioural versus situational questions in a postgraduate admissions multiple mini-interview: a reliability and acceptability comparison. *BMC Med Educ.* 15:75.