
Article:
Webster, Elin S, Paton, Lewis William orcid.org/0000-0002-3328-5634, Crampton, Paul orcid.org/0000-0001-8744-930X et al. (1 more author) (2020) Situational judgement test validity for selection:a systematic review and meta-analysis. Medical Education. pp. 1-15. ISSN 0308-0110

https://doi.org/10.1111/medu.14201

Received: 7 January 2020 | Revised: 20 April 2020 | Accepted: 22 April 2020

DOI: 10.1111/medu.14201

RESEARCH ARTICLE

Situational judgement test validity for selection: A systematic review and meta-analysis

Elin S. Webster[1,2] | Lewis W. Paton[2] | Paul E. S. Crampton[1] | Paul A. Tiffin[1,2]

[1]Hull York Medical School, York, UK

[2]Department of Health Sciences, University of York, York, UK

Correspondence
Paul A. Tiffin, Department of Health Sciences, University of York, Seebohm Rowntree Building, Heslington, York, YO10 5DD, UK.
Email: paul.tiffin@york.ac.uk

Funding information
LWP's research time is part-funded by the UCAT (University Clinical Aptitude Test) Board. PAT is supported in his research by the National Institute for Health Research (NIHR) Career Development Fellowship. This paper presents independent research part-funded by NIHR. The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR or the Department of Health and Social Care.

Abstract

Context: Situational judgement tests (SJTs) are widely used to evaluate 'non-academic' abilities in medical applicants. However, there is a lack of understanding of how their predictive validity may vary across contexts. We conducted a systematic review and meta-analysis to synthesise existing evidence relating to the validity of such tools for predicting outcomes relevant to interpersonal workplace performance.

Methods: Searches were conducted in relevant databases to June 2019. Study quality and risk of bias were assessed using the Quality In Prognosis Studies (QUIPS) tool. Results were pooled using random effects meta-analysis and meta-regressions.

Results: Initially, 470 articles were identified, 218 title or abstracts were reviewed, and 44 full text articles were assessed with 30 studies meeting the final inclusion criteria and were judged, overall, to be at moderate risk of bias. Of these, 26 reported correlation coefficients relating to validity, with a pooled estimate of 0.32 (95% confidence interval 0.26 to 0.39, $P < .0001$). Considerable heterogeneity was observed ($I^2 = 96.5\%$) with the largest validity coefficients tending to be observed for postgraduate, rather than undergraduate, selection studies ($\beta = 0.23$, 0.11 to 0.36, $P < .001$). The correction of validity coefficients for attenuation was also independently associated with larger effects ($\beta = 0.13$, 0.03 to 0.23, $P = .01$). No significant associations with test medium (video vs text format), cross-sectional study design, or period of assessment (one-off vs longer-term) were observed. Where reported, the scores generally demonstrated incremental predictive validity, over and above tests of knowledge and cognitive ability.

Conclusions: The use of SJTs in medical selection is supported by the evidence. The observed trend relating to training stage requires investigation. Further research should focus on developing robust criterion-relevant outcome measures that, ideally, capture interpersonal aspects of typical workplace performance. This will facilitate additional work identifying the optimal place of SJTs within particular selection contexts and further enhancing their effectiveness.

1 | INTRODUCTION

The process of selecting the best candidate for a job is a universal challenge across all industries. There are few such situations where the stakes are higher than when deciding on entrants to medical school. An offer of a place to study is not just an opportunity to gain a university degree, but is usually the gateway to a lifetime career, characterised by both power and responsibility. Ideally, effective medical selection must firstly be 'fair,' in a broad sense.[1] That is, that certain under-represented groups are not unduly disadvantaged by the process; for example, individuals without access to certain resources, such as additional coaching, to help their performance in a specific selection assessment. Second, selection should result in the recruitment of individuals who are both suited to a successful career in the field, and likely to make valuable contributions to society in this regard.[2] These two aims can be seen as complimentary.[1] Though attempts to improve medical selection have often focused on measuring aspects of intellectual ability[3] there is an increasing recognition that 'non-academic abilities' are important when selecting future doctors.[4] Indeed, the majority of disciplinary censures received by practising physicians relates to personal conduct rather than clinical skills or knowledge.[5] However, there are many more challenges to defining and measuring such qualities in contrast to cognitive ability, which can be estimated relatively reliably and validated against academic or educational performance. In this context, we use the term 'non-academic abilities,' in a broad sense, to include qualities or traits relevant to interpersonal functioning, though not directly related to traditional concepts of intelligence, intellectual ability or educational achievement. However, we acknowledge the absence of a single satisfactory label to describe these individual characteristics. Indeed, terms such as 'emotional intelligence' and 'non-cognitive' traits, although sometimes employed, are somewhat contentious.[6,7]

The increasing emphasis on non-academic abilities has led to the rapid development and implementation of assessments intended to evaluate such attributes. In contrast to face to face procedures, such as multiple mini interviews (MMIs), the use of situational judgement tests (SJTs) to measure non-academic abilities is viewed as advantageous, as they are relatively cheap and convenient to deliver at scale.[8] The SJTs, in this context, are an assessment format whereby the test-taker is presented with a series of scenarios depicting an interpersonal situation. The candidate must then usually evaluate several possible behavioural responses to each scenario shown or described. The response format for SJTs varies but commonly involves ranking or rating the potential behaviours in order of either appropriateness or perceived effectiveness. Other response formats also exist, such as the candidate choosing the 'best' and 'worst' behaviours depicted. In the context of personnel selection SJTs can be considered a special kind of procedural knowledge test- that is, they ask a candidate what 'should' be done in response to a portrayed scenario.[9] Consequently, the test-taker either knows what should be done or does not and, by definition, such assessments are not considered prone to 'faking' effects. The procedural knowledge evaluated is assumed a necessary, though not sufficient condition for such behaviours to take place in a similar, actual, workplace situation. This is in contrast to self-report personality measures that are vulnerable to faking in high-stakes testing.[10] Moreover, face to face interviews are also open to different forms of bias, though structuring these processes may reduce some of these influences, to some extent.[11]

It should be emphasised that SJTs are a particular assessment format and this review is concerned only with their use in selecting candidates on non-academic abilities.[12] Although SJTs have been applied in personnel selection for many decades, their popularity increased when they were framed as 'low-fidelity simulations.'[13,14] Such SJTs were conceptualised as employing representations of aspects of actual workplace situations likely to be encountered in the role being applied for. A published meta-analysis of the predictive validity of SJT scores for future workplace performance reported a pooled correlation coefficient of .26.[15] However, to date, there have been no systematic reviews or meta-analytic studies specifically relating to SJTs in medical selection. Given the recent rapid rollout of this approach for evaluating non-academic abilities in medical applicants it seemed timely to conduct such a review.

The SJT test format, in the context of personnel selection, is known to generate scores that are sensitive to a range of design choices, implementation methods and settings. Therefore, a review is also needed to begin to understand which factors are most likely to be associated with the observed validity of the resulting scores from such tools. Such knowledge is essential if SJTs are to be optimally designed, validated and implemented in various stages of medical recruitment. Crucially, the choice of criterion-relevant outcome may be as important as the qualities of the SJT in determining the validity coefficients observed. These are likely to vary according to how feasible they are to obtain, being at least partly dependent on the stage of training being selected into. The traditional approach to SJT development for personnel selection involves creating a series of critical incidents, often based on real-world experience, in order to evaluate how candidates respond, relative to expert consensus. It has been speculated that where such occupational experience is relatively lacking, such as in undergraduate selection, it may be more challenging to develop and validate SJT-based measures for relevant personal qualities.[9] However, key questions remain over the feasibility of using potential alternatives, such as 'construct-driven' SJTs, designed to tap into specific traits, as these may show some of the weaknesses of self-report personality measures.[16] Other factors could also influence the observed validity of SJTs. These may include the medium of delivery (eg, multimedia vs a text-based format) and the choice of outcome criterion against which to validate (eg, self-report vs face to face ratings of aspects of performance).[17] Finally, though it is likely that SJTs generally evaluate 'knowledge of interpersonal effectiveness' the content of such assessments carry a wide variety of labels, such as 'integrity,' 'team working,' 'empathy,' etc. Thus, by observing which outcomes have the strongest relationship with the SJT scores we would hope to gain a greater understanding of the constructs being evaluated in this context. In addition, the Ottawa Consensus

statement on medical selection gave several recommendations for future research, including a call for systematic approaches towards translating evidence into changes in policy and practice.[18]

Thus, the primary aim of this review was to collate the existing evidence for the validity of SJT format assessments used in medical selection for the evaluation of non-academic abilities. A secondary aim was to explore, where possible, the factors associated with the observed validity coefficients, via meta-regressions. Finally, the review was intended to provide both guidance for current practice and, by highlighting existing gaps in the literature, provide an agenda for future research.

2 | METHODS

The protocol for this systematic review was registered prospectively on PROSPERO (CRD42019137761).[19]

2.1 | Selection criteria

Studies were deemed eligible if they investigated any persons undergoing medical selection processes that included an SJT for the evaluation of 'non-academic' abilities. For inclusion, studies had to report on the relationship between SJT scores and an outcome measure that, at least partly, related to non-academic abilities, and was deemed relevant to future or current medical practice. Therefore, relevant outcomes ('validity criteria') would be expected to capture some aspect of interpersonal functioning in the candidate. The outcomes were thus expected to include (though not be limited to): supervisor or tutor ratings; objective structured clinical examination (OSCE) performance, and other ratings of 'integrity' or conscientiousness, or 'success' as a doctor (eg, successful completion of a training stage), etc. No restrictions were placed on study design, though purely qualitative studies were excluded. The inclusion and exclusion criteria are outlined in Table 1.

2.2 | Search strategy and study selection

A search of relevant databases was conducted up until 22nd June 2019. MEDLINE, EMBASE (Excerpta Medica Database), PsycINFO, CINAHL (Cumulative Index to Nursing and Allied Health Literature), ERIC (Educational Resources Information Center), PubMed, MedEdPublish, Scopus, Web of Science and the COCHRANE database were searched. Both keywords and Medical Subject Headings and subject headings were included in the search strategy (see Material S1 and S2, online). Input from a research librarian was sought and appropriate indexing terms were used across all databases.

We identified any available grey literature by searching the University Clinical Aptitude Test (UCAT) Consortium's web page of published research, OpenGrey and Electronic Theses Online Service (EThOS) to identify any material that discussed 'situational judgement tests.' The reference lists of previously conducted, relevant

TABLE 1 The inclusion and exclusion criteria for this systematic review

	Inclusion	Exclusion
Participants	Any persons participating in undergraduate or postgraduate medical selection processes that included a situational judgement test (SJT) as an assessment method Any age, gender, or geographical location	Selection and recruitment in relation to health professions other than medicine
Study design	Observational or trial-based studies Studies involving SJTs being piloted, or implemented in undergraduate level selection for entry to medical school or for postgraduate selection or allocation into further medical training schemes Studies involving some quantitative data collection	Studies where SJTs had been used purely to evaluate applied clinical knowledge only (eg, the Clinical Problem Solving Test used in general practitioner selection) The SJTs not used for selection purposes, such as to support training and development Studies where the SJT scores were used as an outcome, not as a selection method or predictor variable Studies involving only qualitative data collection
Outcome	Outcomes or measures of performance that are likely to be directly or indirectly relevant to interpersonal aspects of current or future medical practise Involve at least some element of evaluation by third parties (eg, supervisors, peers, tutors, colleagues, etc.)	Outcomes based only on self-report measures, such as personality assessments Academic outcomes that do not have a significant interpersonal component to them (eg, those based predominantly on recall of semantic knowledge) Outcomes based only on the participants' own perceptions of their qualities, abilities or performance
Publications	Original, data-based studies published in peer-reviewed journals Publicly available theses from PhD or Master's level degree projects Publicly available reports that have been peer reviewed	Non-empirical literature such as opinion-based articles, editorials and theoretical papers Abstracts from conference proceedings Non-English language

systematic reviews were examined to identify any studies picked up from previous methods. Experts in the field were contacted to confirm and validate our search strategy and results. No restrictions were applied based on date, with all databases being searched from the date of their inception until the end of the review. No restrictions were applied based on publication status, with attempts being made to source unpublished studies, by searching databases that retrieve results of this nature, for example, Web of Science, EMBASE. English language limits were not placed on the searches, but any studies identified for which the full text was not available in English were excluded due to lack of translation facilities. A decision was made to exclude any studies cited only in conference abstracts. This was for several reasons; such abstracts often provided few details on the study and were unlikely to have undergone the same rigour of peer review that studies published in scientific journals underwent. Moreover, it was assumed that studies that were methodologically stronger would be more likely to be subsequently published in the peer-reviewed literature. In addition, our approach included an evaluation of the risk of publication bias. In the event, few relevant conference abstracts were identified and the full texts could not be accessed. Unsuccessful attempts were made to access the full texts via contacting the lead author (ESW) in some cases.

It is also important to note that, for the final set of studies, we only retained those which included at least one outcome that involved some third-party rating of interpersonal functioning that could be considered directly or indirectly related to workplace performance. This deviated from our original registered search protocol, which also included studies using construct-relevant self-report measures as outcomes. This change was made as a result of feedback from peer review of an earlier version of this report. This highlighted that scores derived from self-report measures, such as personality questionnaires, tend to correlate only very modestly, at best, with ratings of workplace performance.[20] However, we observed that excluding the minority of studies that used only self-report measures as outcomes had minimal impact on our key findings.

Two authors (ESW and PESC) were involved in the process of selecting studies for inclusion, independently screening all titles and abstracts identified by the searches. Full text screening was conducted for potentially relevant papers, determining the final studies retained. Disagreements at any stage were resolved through discussion with the other two authors (LWP and PAT) until a consensus was reached. Figure 1 displays the PRISMA (*p*referred *r*eporting *i*tems for *s*ystematic *r*eviews and *m*eta-*a*nalyses) flow diagram outlining the process of study selection. Following this, data extraction was performed by one author (ESW) and checked for accuracy by another (PESC).

2.3 | Quality assessment

To assess study quality, the Quality In Prognosis Studies (QUIPS) tool was used as it is well suited for evaluating the risk of biases (RoBs) seen in predictive and prognostic studies.[21] It was used to rate the studies across six domains of: study participation; attrition; prognostic factor evaluation; outcome measurement; confounding and statistical analysis, and reporting.

A rating for study participation was given for identification, recruitment and description of the participants. The category of 'attrition' looked at whether there were any issues related to dropout or incomplete follow-up and what, if any, attempts were made to correct for these effects. In this review, where applicable, this rating included whether the authors corrected for the possible 'attenuation' effects on observed correlations due to the restriction of range when outcomes are only observed in selected candidates.[22] The SJT score was the 'prognostic factor' evaluated. The rating for this domain was based on how the authors described their methods of SJT content, construct-relevance, design and delivery. With regard to the bias rating of outcome measures, this was based on the description in terms of measurement and potential subjectivity issues and reliability. Confounding looked at whether the studies recorded relevant potential confounding factors and accounted for the influence of these in their analysis. Finally, the domain of statistical analysis and reporting rated to what extent the authors had applied appropriate methods of analysis and the clarity and completeness with which they presented their findings.

The six domains were each given a rating of 'low,' 'moderate' or 'high' RoB. An overall RoB for a study was rated as 'low' if 0 or 1 domains were coded as having moderate to RoB; 'moderate' RoB if 2 or 3 domains were rated in this way, and; 'high' RoB if 4 or more domains were rated as presenting at least a moderate RoB. The overall RoB ratings, along with those domains rated as a moderate to high RoB for each study, are shown the final column of Table S1.

2.4 | Data synthesis

As the literature was assumed to be relatively heterogeneous in nature the results were synthesised narratively.[23] For this review this involved assessing the papers in order to understand the themes underlying the rationale and contexts of the final included studies. The synthesis identified common features of the literature and examined the relative strengths and weaknesses of the findings, and the respective methods on which they were based. The analysis consisted of grouping papers into categories, appraising study quality, and producing a collective synthesis. This information was then summarised formally, as can be seen in Table S1. The narrative synthesis was also used to inform the inferences we drew from the data.

Additionally, validity coefficients (correlations) were pooled using a random effects meta-analysis, allowing for heterogeneity at the study level. Two authors (PAT and LWP) assessed the relevance of the outcomes reported in the identified studies and conferred in order to derive consensus where there was any doubt. As the published papers frequently reported on multiple construct-relevant outcomes relating to the same (or considerably overlapping) study population, we designated these as 'sub-studies.' Therefore, we introduced a

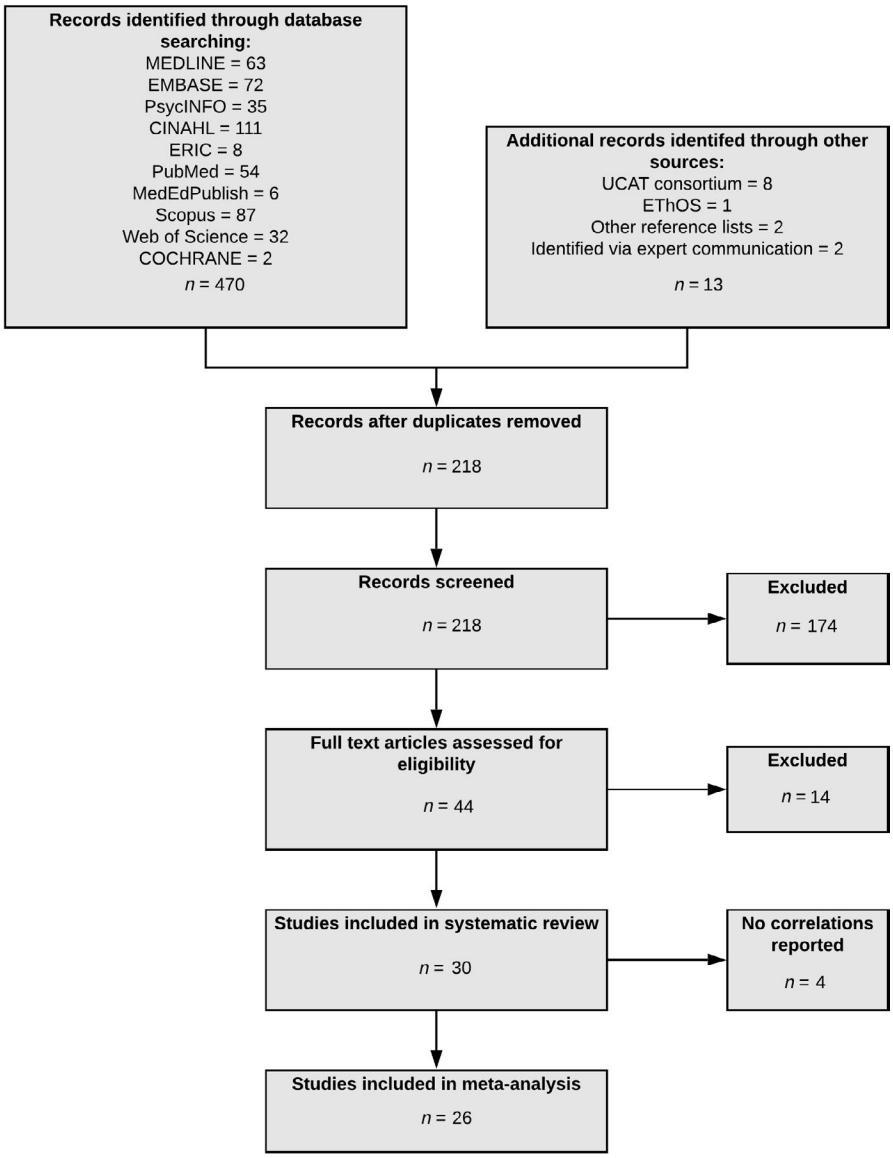

FIGURE 1 PRISMA (*p*referred *r*eporting *i*tems for *s*ystematic *r*eviews and *m*eta-*a*nalyses) flowchart for the systematic review Abbreviations: CINAHL, Cumulative Index to Nursing and Allied Health Literature; EMBASE, Excerpta Medica Database; ERIC, Educational Resources Information Center; EThos, Electronic Theses Online Service; UCAT, University Clinical Aptitude Test

second random effect into the meta-analytic model in order to accommodate the dependency of observations within each shared study population. Thus, there were three-levels (ie involving two random-effects) in our meta-analysis. These levels represented: outcome; population, and paper. We used this model to derive a pooled estimate of the validity of the SJTs used in the relevant studies. Similarly, multi-level meta-regressions were also performed, where applicable, to formally test for any association between sub-study characteristics and the magnitude of the validity coefficients reported. These characteristics formally tested were selected on the basis that: (a) all (or almost all) included studies reported these factors; (b) that there were a sufficient proportion of studies of each type to be likely to observe at least a trend, should it exist, and (c) that there were prior empirical or theoretical reasons to expect some difference in the magnitude of the validity coefficients observed on the basis of the factor. Statistical heterogeneity was assessed using the I^2 statistic.[24] Meta-analysis and regressions were performed in the statistical software R[25] using the metafor package.[26]

3 | RESULTS

In total, the search identified 470 papers, which after removing duplicates left 218 papers to be screened. After title and abstract assessment 174 studies that did not meet the eligibility criteria were excluded. This resulted in 44 full texts to be assessed, of which 30 were found to meet the inclusion criteria and were subsequently retained for analysis in the review (Figure 1). As noted in the published protocol (CRD42019137761),[19] all studies were expected to be observational and this was the case. A total of 10 were cross-sectional studies,[27-35,55] where the outcome was measured at the same time or in the same selection cycle as taking the SJT. A total of 17 were cohort studies[17,36-51] that had a follow-up period before the outcome of interest was measured. Three studies employed a mixture of

cross-sectional and more distal outcomes.[52-54] The length of follow-up across the cohort studies varied from 1 to 9 years after taking the SJT. Full details of the included studies are listed in Table S1.

A total of 11 studies[17,28,29,35,40-43,47,49,50] looked at undergraduate selection for medical school entry, five studies[31,36,37,45,51] at entry to Foundation Year training programmes (the first 2 years of post-qualification training in the United Kingdom [UK]) and 14 studies[27,30,32-34,38,39,44,46,48,52-55] at entry to specialty training. The youngest participant mean age was 17.9 years and the oldest was 34.0 years. Most studies that gave details on sample demographics reported having a majority of female participants. The studies were all conducted in high-income countries, with 19[27-33,36,37,44-49,51-53,55] assessing UK populations, six[17,40-43,50] from mainland Europe, three[35,38,39] from North America, and two[34,54] from Australia. Sample sizes ranged from 51 to 14 131.

A total of 24 studies[27-34,36-40,44-49,51-55] evaluated text-based SJTs, with four[41-43,50] video-based and two[17,35] that included both delivery formats. All but one[35] SJT used a selected response question (SRQ) format. Of these, 12[27,30,31,33,34,36,37,45,48,51,52,54] studies reported on tests that used a ranking-response type format, which involved the test-taker ordering a set of behaviours according to perceived appropriateness or effectiveness; eight studies[17,32,40-43,46,53] asked candidates to choose the most appropriate behaviour/s depicted; six studies[28,29,38,47,49,50] employed a rating scale format, where candidates expressed a judgement about a depicted behaviour. One study[55] used a mixture of formats (rating scale and choosing several appropriate behaviours). One study,[35] which evaluated the Computer-based Assessment for Sampling Personal Characteristics (CASPer) assessment, required open-ended, text-based responses. Two studies[39,44] did not give details of the response formats. A total of 20 studies[17,28-34,36,37,40-43,46,47,50,52-54] included some measure of reliability and internal consistency for the SJT used, which ranged from 0.29 to 0.91. A total of 16 studies[28-32,34,36,40-43,49,50,52,53,55] also provided information in relation to the reliability of the construct-relevant outcome.

3.1 | Outcome measures used

The outcome measures used by the studies to assess the criterion validity of the SJT scores varied but could be approximately divided into two categories. Four studies[43,44,52,53] reported outcomes in both categories:

- *Ratings at face to face, one-off, assessments*: A total of 18 studies[27-35,37,43,44,46,48,52-55] included at least one outcome from face to face, one-off assessments. These included interviews and 'high fidelity' simulations at selection centres and clinical examinations. These assessments were generally considered to capture 'maximal' performance. That is, where test-takers would be expected to put in maximum effort with the aim of achieving as high a score as possible at evaluation in a high-stakes setting.
- *Evaluations of longer term clinical training or work performance*: A total of 16 studies[17,36,38-45,47,49,50-53] employed outcomes related to longer-term evaluations of performance at aspects of future clinical training or workplace performance. For example, supervisor or tutor ratings or Grade Point Averages (GPAs) for aspects of courses with an interpersonal component. Other examples of this type of outcome included issues relating to actual workplace performance, such as recorded lapses of professionalism. It is postulated that such evaluations may be better able to capture 'typical' workplace behaviour, compared to one-off assessments.

3.2 | Risk of bias

Overall, according to the QUIPS tool, the results of the studies were deemed to be at a moderate RoB. The RoB ratings for the studies are summarised in the right hand column of Table S1. A frequent area where the potential for bias was identified was the restriction of range common, by definition, to selection studies, whereby outcomes were only observed for those recruited. Not all studies attempted to correct for these, and other attenuating effects, such as imperfect reliability in the SJT or outcome. Other common potential sources of bias were unreported or relatively poor reliability (<0.7) of either the SJT used or the outcome measure. Moreover, some studies did not provide adequate descriptions of the population, SJT or outcome characteristics. Nevertheless, it should be noted that the potential sources of bias identified would tend to lead to a systematic underestimation of the relationship between the SJT scores and the construct of interest. Thus, the results of the studies could be considered likely to be relatively conservative, especially those at higher overall RoB.

Concerning potential 'confounding,' we focused on whether the influence of academic performance had been controlled for. That is, whether the scores from the SJTs were likely to add incremental value above and beyond the traditional measures of academic or intellectual ability that are already widely employed. A total of 17 studies[17, 27, 30-32,37,38, 40-43,45,46,48,51,52,55] used a measure of cognitive or academic ability alongside the SJT and thus received a low RoB rating in this domain. For the purposes of this review, we did not consider demographic variables such as age and sex as potential confounders, as they are largely not used in medical selection.

3.3 | Meta-analysis

Overall, across our 30 included studies, all but four studies[39,44,45,51] reported correlation coefficients between SJT scores and construct-relevant outcomes. In total, the remaining 26 published studies reported on 50 outcomes, which, for the purposes of the meta-analysis, were treated as separate studies (sub-studies). Thus, the second random effect used in our meta-analytic model accommodated potential dependency in outcomes related to these sub-studies and related to 24 populations. Consequently, the three-level random-effects meta-analysis estimated the pooled correlation across all 50 sub-studies to

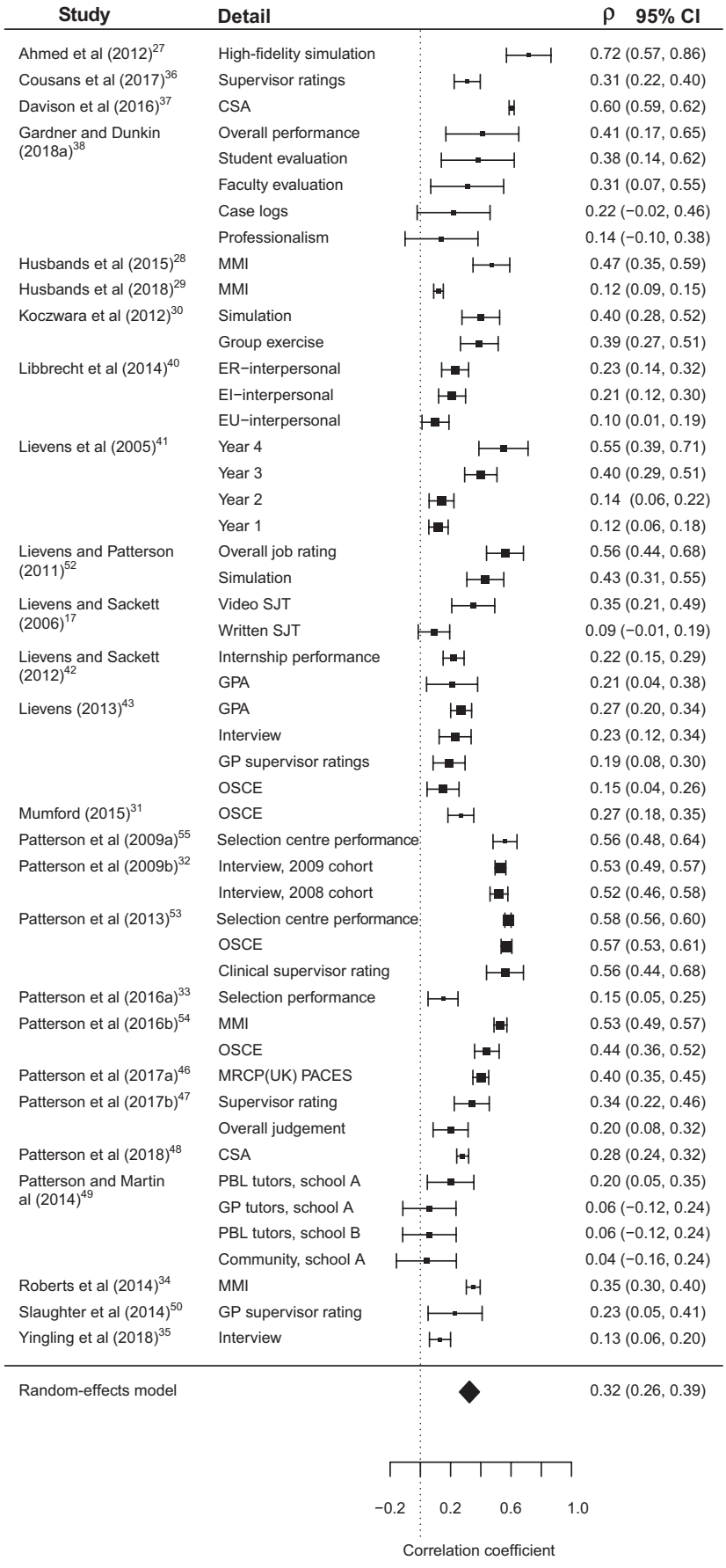

FIGURE 2 Forest plot depicting the relative effect sizes for the 50 validity coefficients reported by 26 studies. ρ = correlation

be 0.32 (95% confidence interval [CI] 0.26 to 0.39, $P < .0001$). The results are summarised in the forest plot shown in Figure 2. Substantial heterogeneity was observed as indexed by an I^2 statistic of 96.5%. The I^2 value reflects the percentage of variation in the results across studies that is due to heterogeneity in studies (eg, different designs, outcomes and populations, etc.) rather than chance.[56] In this case the I^2 statistic was close to 100.0% suggesting only a small proportion of variation was due to chance alone.

There are many design features in an SJT validation study that may influence the magnitude of the validation coefficient observed.[57] However, in the present case there were a limited number of such characteristics that could be formally statistically evaluated. This was because such features had to be explicitly described in almost all the studies, with sufficient variation across them to plausibly evaluate for the presence of any trends. Almost all studies reported on whether text or video was used to present the scenarios (ie, the stimulus), the setting (undergraduate vs postgraduate), whether the outcome was longitudinal in nature (ie, captured more than a year after the SJT was administered) as opposed to cross-sectional, if the outcome was captured on a one-off occasion rather than via a more prolonged period of assessment, and whether or not any correction for attenuation of the observed correlation was applied. There were also a priori reasons for hypothesising that these factors might be related to the reporting of higher or lower values for the validity coefficients. Therefore, the potential associations between these factors and the magnitude of the validity coefficients were formally tested using meta-regression analyses. Both univariable and multivariable models were tested in this regard. The results are shown in Table 2.

As can be seen from Table 2, only study setting was a statistically significant univariable predictor of reported correlation coefficient, with studies of postgraduate medical selection tending to report larger validity coefficients compared to those conducted in undergraduate setting ($\beta = 0.21$, 95% CI 0.11 to 0.31, $P < .001$). This remained a significant predictor of the magnitude of the correlation coefficient when including the other study characteristic variables in a multivariable meta-regression ($\beta = 0.23$, 95% CI 0.11 to 0.36, $P < .001$). This inferred, that, on average, when controlling for the influence of the other variables included in the model, postgraduate-based studies reported regression coefficients 0.23 larger in magnitude compared to those from undergraduate settings. Interestingly there was no univariable relationship between the magnitude of the reported correlation coefficients and whether attenuation effects were corrected for or not. However, when the potential influence of the other factors were controlled for within a multivariable model this association became statistically significant ($\beta = 0.13$, 0.03 to 0.23, $P = .01$). No significant independent trends were observed relating to cross-sectional vs longitudinal outcomes, the use of video vs text SJT format, or the use of one-off (vs longer term) assessments as outcomes.

Figure 3 displays the funnel plot for this meta-analysis.[58] This depicts the distribution of the magnitude of the validity coefficients (ie, effect size) on the horizontal axis against the standard error (study precision) on the vertical axis. The latter is related to study size and hence, power. Funnel plots are used to help evaluate both heterogeneity and the risk of publication bias. Regarding heterogeneity; if differences between the findings of studies were due purely to sampling error (ie, evaluating the SJTs in similar, randomly selected populations of test-takers) then most, if not all the point estimates would fall within the pale triangle, representing the 95% CIs. As can be seen from Figure 3, numerous point estimates fall outside this area, suggesting considerable heterogeneity in terms of study design and/or population characteristics. Marked asymmetry within funnel plots are consistent with the possibility of publication bias. That is, small studies, which report modest or negligible effect sizes may be less likely to be published than ones, which may show relatively large validity coefficients. Studies with relatively few participants may lack the power to detect the true underlying relationship between a predictor and an outcome. Therefore both small and large effect sizes may be more likely to be due to chance, though the latter may be more likely to result in a study being published. In a funnel plot, this bias can manifest as asymmetry with a relative paucity of studies in the lower left quadrant of the chart. This may provide evidence that there are fewer published small studies reporting modest effect sizes than may be expected by chance. As can be seen in Figure 3, the funnel plot is relatively symmetrical in this respect, providing no indication of publication bias.

4 | DISCUSSION

This is the first review to systematically collate and synthesise evidence relating to the use of SJT-format assessments for selecting candidates into medical training based on non-academic abilities.

TABLE 2 Results from the univariable and multivariable meta-regression analyses

	Univariable (raw) coefficients		Multivariable (adjusted) coefficients	
Study characteristic	**β (95% confidence interval)**	***P*-value**	**β (95% confidence interval)**	***P*-value**
Postgraduate setting (vs undergraduate setting)	0.21 (0.11 to 0.31)	<.001	0.23 (0.11 to 0.36)	<.001
Cross-sectional design (vs longitudinal)	0.06 (−0.02 to 0.15)	.17	0.05 (−0.05 to 0.14)	.34
Video-based SJT (vs text-based SJT)	−0.06 (−0.20 to 0.07)	.36	0.03 (−0.10 to 0.16)	.61
One-off assessment (vs longer term evaluation)	0.07 (−0.02 to 0.17)	.13	0.02 (−0.08 to 0.13)	.70
Corrected for attenuation	0.08 (−0.04 to 0.20)	.18	0.13 (0.03 to 0.23)	.01

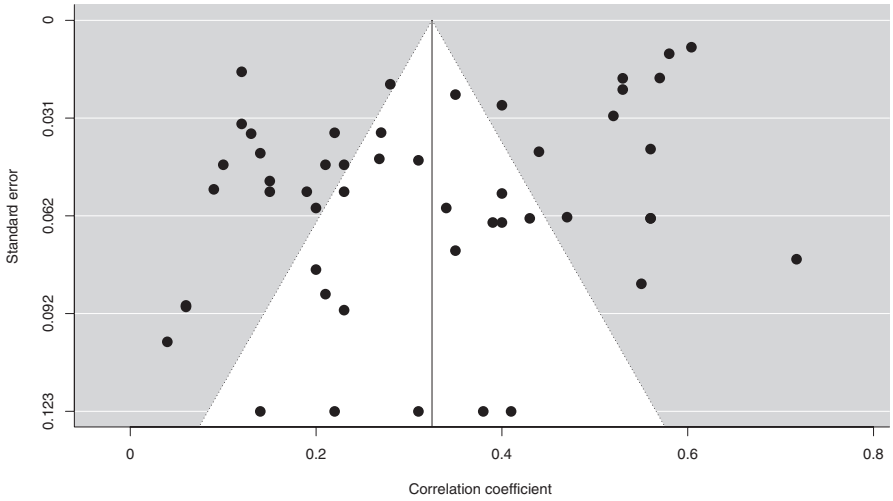

FIGURE 3 Funnel plot for a random-effects meta-analysis of results from 26 studies, which report correlation values relating to 50 outcomes (sub-studies)

We identified a substantial number of studies that reported evidence in relation to the effectiveness of such tools to select for aspects of interpersonal effectiveness. Overall, importantly the findings suggest that the scores derived from SJTs correlate, at least modestly, with metrics of performance on assessments that may assess non-academic, interpersonal abilities. The majority of studies observed statistically significant ($P < .05$) associations between test scores and construct-relevant outcomes. We noted that the relationship between test scores and outcomes appeared to be strongest for studies conducted in postgraduate medical selection settings. On multivariable, though not univariable, analysis there was a trend noted for correction for attenuation to be independently associated with slightly higher average observed validity coefficients. There were no statistically significant trends observed for studies using cross-sectional, rather than longitudinal outcomes, to report higher validity coefficients. Likewise, there were no statistically significant trends for studies using one-off assessments to cite higher coefficients compared to ones that relied on longer term evaluations. No associations between the media type used for the SJT and the magnitude of the validity coefficients were observed.

Overall, the pooled correlation estimate was .32. This is relatively close to, though slightly higher than, the average of .26 for criterion related coefficients reported in a previous meta-analysis of studies of SJTs for personnel selection generally.[15] Initially this might seem to infer that the predictive abilities of such tools are relatively modest. However, this finding must be placed in the overall context of medical selection and the other assessment types frequently employed. Here, candidates tend to be relatively homogenous and highly performing, making it more difficult to demonstrate clear differences between individuals, both on testing and in terms of outcomes. Moreover, the usual problems with restriction of range in selection settings apply, in that unsuccessful candidates tend not to have relevant outcomes captured. Furthermore, the imperfect reliability of predictors and outcomes further attenuate the observed relationship between the two.[59] Also, measuring non-academic personal qualities is undoubtedly more challenging than attempting to capture cognitive ability, where there has been historically a great deal of experience and the traits under assessment are well defined.[6] Moreover, the predictor-outcome coefficients we observed in the present review are frequently similar in magnitude, and sometimes larger, compared to those cited for cognitive assessments used to evaluate problem-solving ability in medical school applicants.[59-62]

There are numerous possible reasons that may explain why the validity coefficients for studies set in undergraduate vs postgraduate settings were, on average, lower. These explanations are not mutually exclusive. First, it may be particularly challenging to obtain appropriate, construct-relevant outcomes for validation studies in relation to undergraduate selection. For example, colleagues or supervisors of practising doctors will have substantial opportunities to observe and rate workplace performance. In contrast, tutors or supervisors of students may have more limited contact, often outside of a clinical setting, on which to base such evaluation of interpersonal effectiveness. It could be argued that, for undergraduate selection studies, distal outcomes may be more relevant to actual job performance. However, traits or abilities may change over time, with maturity and/or training. Thus, the time lag between the selection test and outcome may actually attenuate the observed correlation between the two. This attenuation may also be exacerbated if candidates achieving only low scores on an SJT are not admitted to medical school, substantially restricting the range of performance that can be observed in both predictor and outcome variables. However, we noted that the association between coefficient magnitude and undergraduate vs postgraduate setting was independent of other factors in our multivariable model. Nevertheless, given the relatively small number of studies in the analysis, this finding does not preclude choice of outcome playing a role in explaining this observation. Developing SJTs for personnel selection traditionally depends on eliciting 'critical incidents' based on actual or plausible workplace situations. Such incidents are also often likely to tap into cognitive abilities and, perhaps to an extent, clinical judgement, which will increase their association with actual job performance. In contrast, when selecting into early stages of training there will be fewer relevant scenarios to sample from as candidates may have had little

or no relevant workplace exposure. Moreover, as work experience accumulates it may cause traits or knowledge to coalesce in individuals, rendering them more easily measurable, but also increasing the association with actual job performance. Such potential factors have led to speculation that this traditional approach to SJT design may be less effective for selection into early, compared to later, career stages of medicine.[9] However, such a conclusion should not be drawn from the current findings, given the practical challenges highlighted earlier regarding establishing validity. It should also be noted that the use of SJTs for selecting candidates into undergraduate medical studies is a relatively new development. Thus, it may be that with additional experience, the properties of such instruments continue to improve in this setting, resulting in validity coefficients comparable to those seen, on average, in postgraduate settings.

In contrast to a previous meta-analysis of SJTs we did not observe an obvious trend for video-based SJTs to report higher validity coefficients compared to text-based formats.[63] However, only six studies[17,35,41-43,50] including video-based format SJTs were identified and the heterogeneity in selection settings and outcomes used may have obscured such a trend, were it to exist in this context. Interestingly, we observed a multivariable, though not univariable, association between correction for attenuation in the studies and the validity coefficients reported. On average, by definition, one would expect the corrected coefficients in selection studies to be larger. It is therefore likely that, given the relatively small number and heterogeneity of the studies identified, the relatively modest potential influence of this factor may have been initially obscured, at univariable analysis, by other design features.

Over half (n = 17) of the studies[17,27,30-32,37,38,40-43,45,46,48,51,52,55] identified attempted to estimate the incremental validity of the SJT being evaluated, above and beyond other selection assessments, such as tests of cognitive ability, academic performance and clinical knowledge. Of those that did, only two[31,45] did not report any evidence of incremental validity. However, most of the studies that did demonstrate some incremental validity for the SJT scores reported relatively modest values, often in the range of approximately 5.0 to 10.0% of additional variance in the outcome accounted for. Again, given the particular challenges outlined earlier in establishing the construct-validity of selection measures in medicine, such modest values for incremental validity are understandable.[59] Moreover, incremental validity is likely to be greatest when adjusting for the effects of constructs that are different from those evaluated by the SJTs. Such a trend was not obvious in the results of the final pool of studies (see Table S1), though other design issues, such as the choice of outcome measure, are also likely to have played a substantial role in determining the degree of incremental validity observed.

4.1 | Strengths and potential limitations

We used a rigorous systematic review process, with a prospectively registered search strategy, which identified a substantial number of primary studies for inclusion. These strengths aside, there were several potential limitations to the conduct of the review. First, it may be that we failed to identify unpublished studies, which observed weak or absent correlations between SJT scores and an outcome of interest (ie, publication bias). The review also excluded studies that only reported their findings in conference abstracts. However, we note that our estimated average reported validity coefficient of 0.32 is close to that reported by a previous meta-analysis of SJTs for personnel selection that did include unpublished studies. A previous similar meta-analysis that only included published studies reported a higher pooled validity coefficient of 0.34.[64] Moreover, our funnel plot (Figure 3) did not provide evidence of publication bias. Due to the lack of translation facilities the studies were restricted to those published in the English language. Nevertheless, only one study was excluded for this reason.[65] Not all of the final study results could be entered into the meta-analysis as some did not cite a correlation coefficient as validity evidence, having employed categorical outcomes. However, only four studies[39,44,45,51] were excluded from the meta-analysis on these grounds.

As with any systematic review, the primary limitation was the quality of the studies included. Overall, the studies were rated as at moderate RoB. A number of the studies were characterised by relatively small sample sizes or high rates of participant attrition. Several of the included studies[32,46,52,53] involved the same population of participants, which followed up the original participants or re-analysed the data. However, our use of a second random effect, to adjust for shared populations, where different or follow-up outcomes were reported, should have adjusted for this potential source of bias. In addition, given the relatively high levels of heterogeneity in the final pool of studies identified, as indicated by the I^2 statistic and the degree of scatter on the funnel plot (Figure 3), some caution must be exercised when drawing inferences about particular factors influencing the observed validity of the SJTs. This is because the differing results observed are highly unlikely to be due to random variation in sampling alone. Some of these differences will be explained by aspects of study design, such as context and the outcomes selected, which were captured in our meta-regression. However, inevitably, there will have been other factors, which would have either not been reported consistently in the studies, or captured as part of our data extraction and analysis. Ideally all the design features that may have been relevant to criterion-validity would have been formally tested for their influence on the results. However, due to the number and nature of the studies identified only five of these factors were formally evaluated in the meta-regression. Moreover, given the relatively small numbers of each study type these tests may have been underpowered, and so some caution must be exercised in interpreting the results.

4.2 | Implications for policy and practice

The use of the SJT format to evaluate non-academic attributes is becoming increasingly widespread across medical selection and the results of this review would support their general validity in this context. The majority of the studies reported moderate, rather

than large, predictor-outcome correlation coefficients. However, as highlighted earlier, these are comparable to those frequently cited for other widely accepted medical selection tools. Having established that SJTs generally have both predictive, as well as incremental, validity in this context there is a question about their optimal place within the selection process. A previous review of the evidence for personnel selection approaches in medicine suggested that SJTs, along with MMIs, (cognitive) aptitude tests, academic record and selection centres, were fairer and more effective than personal statements, references and traditional interviews.[66] Consequently these latter three, relatively unreliable, methods of evaluating personal qualities in medical applicants are less preferable compared to SJTs and the structured observations employed by MMIs and selection centres. Structured interviews, such as MMIs, seem to demonstrate acceptable reliability and validity if implemented appropriately.[67] However, they are relatively resource intensive compared to SJT format assessments. The SJTs used in personnel selection are generally experienced as relatively easy tests.[68] Therefore, they tend to discriminate most accurately between relatively poorly performing test-takers. This implies they may best serve as cost-effective 'screen outs' at an early selection stage when considering, which applicants to progress to face to face interview processes. Our review also highlighted that, where evaluated, SJT scores generally demonstrate some degree of incremental validity over and above tests of cognitive ability or clinical knowledge. Therefore, the use of SJTs in combination with such measures as an early stage of selection into undergraduate or postgraduate training posts seems justified. Indeed, for the studies that reported on SJTs already implemented in selection, this was often how the tests were described as being used. Moreover, the positive correlations reported between SJT and face to face assessment scores reported by relevant studies would provide additional evidence for their usefulness as a screening tool for selection for interview. Conversely, it has also been highlighted that, in some contexts, the use of resource intensive face to face processes add little incremental predictive power over and above a cheaper battery of written assessments that include SJTs. For example, a review of the selection system into the UK general practitioner (GP) training selection scheme evaluated the incremental validity of scores from a selection centre (with simulated consultations) over and above those derived from an earlier battery of written tests, which included a clinical knowledge test and an SJT. The authors reported that the selection centre scores predicted only an additional 3.0 to 4.0% of variance in the later clinical skills assessment exam, taken as part of subsequent specialty training.[37] Similar findings were reported in relation to selection into Australian GP training when comparing an MMI to an SJT.[54] Such findings led the authors of the former report to speculate that, given the high costs associated with vacant GP posts, it may be more cost-effective to dispense altogether with the face to face stage of selection in this context.[37] There may also be other circumstances in which selectors may wish to only interview candidates obtaining a mid-range score on an SJT. For example, this approach could possibly be justified where the numbers of both applicants and places are relatively large compared to the available resources to perform face to face assessments and those performing well at an SJT are known to be at minimal risk of receiving poor interview-based ratings. However, though such choices may be justified they would have to be based on some preliminary evidence; the cost-effectiveness of SJTs in conjunction with more resource-intensive selection stages may assumed to be reasonably sensitive to context.

As highlighted earlier, the low to moderate validity coefficients reported for most SJT evaluations are comparable with those cited for cognitive (problem-solving) ability to predict medical academic performance. Thus, it could be justifiable that a similar weight be placed on SJT performance as on the latter assessment scores. However, some caution in this regard should be exercised. In general, due to their measurement properties and precision, the cognitive assessments used in medical selection are generally able to differentiate candidates even at the upper end of ability.[69] This is less true of SJTs, which tend to be superior at discriminating between average to low performers, hence their suitability to be used as early stage screening assessments.[68] Indeed the weight placed on SJTs, relative to other measures, in medical selection has been debated in relation the allocation process for UK medical graduates to be placed on the country's Foundation training programme (the first 2 years of postgraduate medical training).[70,71] In this case equal weight is currently given to the scores from a 2.4 hour long SJT and the education percentile measure (EPM) derived from academic performance in the previous 5 to 6 years of medical school.[72] Nevertheless, it should be highlighted that, in this situation, the focus of discussion has not been with the quality of the specific SJT used, but rather the relative weight placed on the scores.[73] Indeed, the authors of the validity study for the allocation process for the Foundation programme found evidence for the effectiveness of the SJT used in this context, though suggested that a relatively reduced weighting be placed on the SJT performance in this situation.[51]

To summarise, the current state of evidence supports the use of such SJTs within medical selection, usually in conjunction with tests of knowledge and/or cognitive ability. Such assessments may best serve as a way of deciding, which candidates should progress to more resource intensive assessment processes. However, in some circumstances it may be defensible, and more cost-effective, to limit face to face processes to those applicants that score in a certain (middle range) or dispense with such a final stage of selection altogether. Such situations may include those where there is a low applicant to vacancy ratio and a strong imperative to fill training places.

4.3 | Directions for future research

Collating and summarising the empirical evidence to date in this area enables us to describe a clear agenda for future research in the field. First, the identified studies used a wide variety of outcomes in order

to obtain evidence to support the validity of the SJTs. Many of these would have been expected to tap into a whole range of traits and abilities that might be presumed to be relevant to performance in a health care setting. However, ideally such outcomes would be more explicit in terms of the constructs that they were tapping into, and indeed their precision (reliability) in terms of their measurement ability. The importance of matching a test score to a criterion-relevant outcome has been previously emphasised.[41] Moreover, the development of a framework for outcome criteria in medical selection has previously been highlighted as a research priority in the field.[66] This should lead to the development of robust construct-relevant measures that can be used in SJT validation studies. In the field of emotional intelligence (EI) research there have been recent attempts to develop taxonomies of situations that tap into different traits and abilities.[74] It may be that such classification systems can be used to base the development of novel instruments that are able to measure various aspects of interpersonal functioning in test-takers applying for health care training or roles. In this regard, particular thought should be given to whether such outcome-measurement approaches will mainly capture typical or maximal interpersonal performance. Indeed, the EI literature sometimes classifies measures as either 'trait-' or 'ability-based' depending on whether they are aiming to evaluate typical or maximal performance in relation to social and emotional functioning.[75] High-stakes tests, including SJTs, might be expected to capture maximal performance more effectively than typical performance. However, for employers (and patients) it would be typical performance that would be of most concern. Nevertheless, creating measures that effectively capture typical performance in high-stakes situations is challenging. Possibilities for creating such assessment approaches may include use of more immersive formats, such as virtual reality, or (more ethically fraught) the use of misdirection or deception to detect dishonest behaviour, such as the tendency to fake test responses.[76] Though it is likely that such approaches would be relatively resource intensive, such novel measures could be used to validate SJTs that could then be deployed cheaply, at scale. Moreover, such outcomes could be used in a cross-sectional way, avoiding the delay involved in obtaining longitudinal outcomes. The most pressing need for robust outcome measures would seem to be in relation to early stages of career selection, where observations of actual workplace performance may be scant or unavailable.

Second, though most studies, which investigated the incremental validity of SJTs reported some evidence to support this, the magnitude and nature of this varied. Consequently, understanding the degree to which SJT scores add value, in which medical selection contexts, is a priority if they are to find their most appropriate place within health care recruitment. Indeed, it has been highlighted that understanding the optimum combination and weighting of selection tools requires further research.[66] In this regard, a mathematical model for personnel selection has been proposed, based on 'pareto-optimality.'[77] In this framework, the progress towards one aim can be furthered although not negatively impacting another. Thus, there are both more and less efficient ways of arranging selection systems so that the candidates with the most potential are chosen at the same time minimising the adverse impact on under-represented groups. Indeed, there have been suggestions that placing more weight on SJT performance, especially earlier in the selection process, could facilitate widening access to medicine.[78,79] Applying such a pareto-optimal framework may help to indicate the most efficient use of SJTs in various medical selection situations, modelling the likely impact on the population chosen.

Third, there are some key questions in relation to design issues. For example, we identified relatively few studies using video or multimedia format to present situations. It is thus, currently unclear whether providing more immersive experiences to test-takers increases the predictive validity of such tests. Almost all the SJTs in this review used some kind of SRQ response format. Thus, it is not possible to say whether removing such answering cues, for example, by using free text responses, would increase the validity of such assessments. Indeed, in this regard, there is some evidence that, at least for semantic knowledge tests, free text response format questions are generally experienced as more difficult compared to the equivalent SRQs.[80] With advances in natural language processing and machine learning comes the increasingly plausible possibility of automating, or semi-automating, the scoring of such responses. Machine learning may also offer the possibility of side-stepping issues with developing SJT scoring keys, though this would be dependent on the availability of robust outcomes against which to train such systems.[81] The use of more engaging and immersive technologies, using augmented and virtual reality[82] or gamification[83] could also be harnessed in a way, which makes it more feasible to capture more typical (rather than maximal) interpersonal performance, even in a high-stakes selection setting. Consequently, it may be possible to develop and enhance the SJT format in a way, which renders them more effective, though still able to be delivered at scale. It has also previously been suggested that a more construct-driven approach, where particular traits are targeted, may be useful to SJT development in some circumstances. In particular this framework has been raised as a possibility for SJTs used in selection situations where there are relatively few workplace situations to sample, and relatively little on the job experience for test-takers to draw on.[9] However, it is currently unknown if such tests, which, psychometrically, often behave more like traditional personality measures, would be valid in high-stakes situations, where faking effects may come into play.[9,16]

Fourth, modelling or quantifying the impact of these selection tools on the actual demographics, and indeed the effectiveness, of the medical profession is required. In this regard, numerical simulation methods, such as those previously applied to medical selection situations, may be useful.[5,60] As the widespread use of SJTs in undergraduate medical selection is a relatively recent development, there is an opportunity to evaluate the footprint of such policy changes over the near future. This could be performed via tracked cohorts, such as those whose information is captured by the UK Medical Education Database.[84] If such SJTs are successful in their aims then a footprint should be observed in terms of improved patient satisfaction and health outcomes. Similarly rates of complaints and professionalism breaches should decline. In order to control for such effects being obscured by other secular trends it may be necessary to employ quasi-experimental designs, or causal inference approaches to analyses

of relevant observational data. It should also be possible to provide estimates of the likely impact of both the nature of the tests, and the manner in which they are implemented, on access to medicine to traditionally under-represented groups, such as those from ethnic minorities or less advantaged socio-economic backgrounds.

More generally, it should be highlighted that all the included studies were from high-income countries. Previous research has found that SJT methodology is typically transportable for use in recruitment settings in other countries.[85] However, there is a need for further research relating to the validity and impact of such tools across diverse settings and cultures. We also noted that most of the published research studies identified were led by, or involved, test developers. Therefore, some degree of conflict of interest would be present. It would therefore be desirable, where possible, for a greater number of independent evaluations to be conducted in the future. This may be challenging as much of the testing expertise currently lies within the commercial sector. It is also the case that when SJTs need to be deployed at scale commercial organisations generally have to be involved. Therefore, it is likely that evaluating the effectiveness of widely used SJTs would involve some degree of partnership with industry. Moreover, academics are themselves not free of potential conflicts of interest.

5 | CONCLUSIONS

Our findings suggest that SJTs used for evaluating non-academic abilities in medical selection generally demonstrate moderate predictive validity for construct-relevant outcomes. Thus, SJTs are likely to be useful as part of a well-designed selection system, most probably at an early stage of recruitment, to help support decision-making about progressing to more resource intensive assessments. Further research should focus on understanding the underlying reasons for the relatively lower validity coefficients observed in undergraduate settings. This should include the development of robust, cross-sectional, construct-relevant outcomes suitable for use in earlier medical career stages. Additional work should establish the incremental validity and cost-effectiveness of such tools in differing contexts, and thus, their optimum place within the selection process. This will also help further our understanding of the most effective ways of evaluating personal qualities in this group.

AUTHOR CONTRIBUTIONS

PAT conceptualised the review. ESW prepared the first draft of this paper. ESW also led on data acquisition, with substantial contributions from PESC. LWP and PAT led on data synthesis. All authors (ESW, LWP, PESC and PAT) contributed to the development of the protocol, interpretation of findings, the critical appraisal and revision of the document, as well as approved the final manuscript for submission, and agreed to be accountable for all aspects of the work.

ACKNOWLEDGEMENTS

None.

CONFLICTS OF INTEREST

The UCAT board pay for a portion of LWP's research time, and LWP has received travel expenses for attendance at a UCAT Consortium meeting. PAT has previously received research funding from the Economic and Social Research Council, the Engineering and Physical Sciences Research Council, the Department of Health for England, the UCAT Board, and the General Medical Council. In addition, PAT has previously performed consultancy work on behalf of his employing university for the UCAT Board and Work Psychology Group and has received travel and subsistence expenses for attendance at the UCAT Research Group.

ETHICAL APPROVAL

Not applicable.

ORCID

Elin S. Webster https://orcid.org/0000-0001-8478-994X
Lewis W. Paton https://orcid.org/0000-0002-3328-5634
Paul E. S. Crampton https://orcid.org/0000-0001-8744-930X
Paul A. Tiffin https://orcid.org/0000-0003-1770-5034

REFERENCES

1. Tiffin PA, Alexander K, Cleland J. When I say … fairness in selection. *Med Educ*. 2018;52(12):1225-1227.
2. Cleland J, Dowell J, McLachlan JC, Nicholson S, Patterson F. *Identifying Best Practice in the Selection of Medical Students*. London, UK: General Medical Council; 2012.
3. Kelly M, Tiffin PA, Mwandigha LM. Aptitude testing in healthcare selection. In: Patterson F, Zibarras L, eds. *Selection and Recruitment in the Healthcare Professions*. London, UK: Palgrave; 2018:27-50.
4. Emanuel EJ, Gudbranson E. Does medicine overemphasize IQ? *JAMA*. 2018;319(7):651-652.
5. Tiffin PA, Paton LW, Mwandigha LM, McLachlan JC, Illing J. Predicting fitness to practise events in international medical graduates who registered as UK doctors via the Professional and Linguistic Assessments Board (PLAB) system: a national cohort study. *BMC Med*. 2017;15(1):66.
6. Tiffin PA, Paton LW. When I say … emotional intelligence. *Med Educ*. 2020;https://doi.org/10.1111/medu.14160
7. Petrides KV. Psychometric properties of the trait emotional intelligence questionnaire (TEIQue). In: Stough C, Saklofske DH, Parker JDA, eds. *Assessing Emotional Intelligence: Theory, Research and Applications*. New York, NY: Springer; 2009:85-101.
8. Patterson F, Zibarras L, Ashworth V. Ashworth V. Situational judgement tests in medical education and training: research, theory and practice: AMEE Guide No. 100. *Med Teach*. 2016;38(1):3-17.
9. Tiffin PA, Paton LW, O'Mara D, MacCann C, Lang JWB, Lievens F. Situational judgement tests for selection: traditional vs construct-driven approaches. *Med Educ*. 2020;54(2):105-115.
10. McFarland LA, Ryan AM. Variance in faking across noncognitive measures. *J Appl Psychol*. 2000;85(5):812-821.
11. Alonso P, Moscoso S. Structured behavioral and conventional interviews: Differences and biases in interviewer ratings. *J Work Organ Psychol*. 2017;33(3):183-191.
12. Pearce J, Jackel B. SJT, MCQ, ETC… The worrying conflation of format and content. *Med Educ*. 2018;52(9):993-993.
13. Lievens F, Peeters H, Schollaert E. Situational judgment tests: a review of recent research. *Pers Rev*. 2008;37(4):426-441.
14. Motowidlo SJ, Dunnette MD, Carter GW. An alternative selection procedure: the low-fidelity simulation. *J Appl Psychol*. 1990;75(6):640-647.

15. McDaniel MA, Hartman NS, Whetzel DL, Grubb WL. Situational judgment tests, response instructions, and validity: a meta-analysis. *Pers Psychol*. 2007;60(1):63-91.
16. Lievens F. Construct-driven SJTs: toward an agenda for future research. *Int J Test*. 2017;17(3):269-276.
17. Lievens F, Sackett PR. Video-based versus written situational judgment tests: a comparison in terms of predictive validity. *J Appl Psychol*. 2006;91(5):1181-1188.
18. Patterson F, Roberts C, Hanson MD, et al. 2018 Ottawa consensus statement: selection and recruitment to the healthcare professions. *Med Teach*. 2018;40(11):1091-1101.
19. Webster E, Tiffin PA, Paton LW, Crampton P. The predictive validity of situational judgement tests (SJTs) in medical selection: a systematic review. PROSPERO: International prospective register of systematic reviews. NIHR; 2019. https://www.crd.york.ac.uk/PROSPERO/display_record.php?RecordID=137761. Accessed December 3 2019.
20. Hurtz GM, Donovan JJ. Personality and job performance: the Big Five revisited. *J Appl Psychol*. 2000;85(6):869-879.
21. Hayden JA, van der Windt DA, Cartwright JL, Côté P, Bombardier C. Assessing bias in studies of prognostic factors. *Ann Intern Med*. 2013;158(4):280-286.
22. Zimmermann S, Klusmann D, Hampe W. Correcting the predictive validity of a selection test for the effect of indirect range restriction. *BMC Med Educ*. 2017;17:246.
23. Lucas PJ, Baird J, Arai L, Law C, Roberts HM. Worked examples of alternative methods for the synthesis of qualitative and quantitative research in systematic reviews. *BMC Med Res Methodol*. 2007;7(1):4.
24. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327(7414):557-560.
25. R Core Team. *A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing;2020.
26. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36(3):1-48.
27. Ahmed H, Rhydderch M, Matthews P. Can knowledge tests and situational judgement tests predict selection centre performance? *Med Educ*. 2012;46(8):777–784.
28. Husbands A, Rodgerson MJ, Dowell J, Patterson F. Evaluating the validity of an integrity-based situational judgement test for medical school admissions. *BMC Med Educ*. 2015;15:144.
29. Husbands A, Dowell J, Homer M, McAndrew R, Greatrix R.Exploring the relationship between the UKCAT situational judgement test and the multiple mini interview. 2018. Published by the UCAT Board. https://www.ucat.ac.uk/media/1277/ukcat-sjt-mmi-report-march-2018.pdf.
30. Koczwara A, Patterson F, Zibarras L, Kerrin M, Irish B, Wilkinson M. Evaluating cognitive ability, knowledge tests and situational judgement tests for postgraduate selection. *Med Educ*. 2012;46:399–408.
31. Mumford S.The situational judgment test: cognition, constructs and criterion validity, PhD Thesis, University of Sheffield. 2015. http://etheses.whiterose.ac.uk/9597/
32. Patterson F, Carr V, Zibarras L, et al. New machine-marked tests for selection into core medical training: evidence from two validation studies. *Clin Med*. 2009;9(5):417–420.
33. Patterson F, Knight A, McKnight L, Booth TC. Evaluation of two selection tests for recruitment into radiology specialty training. *BMC Med Educ*. 2016;16:170.
34. Roberts C, Clark T, Burgess A, Frommer M, Grant M, Mossman K. The validity of a behavioural multiple-mini-interview within an assessment centre for selection into specialty training. *BMC Med Educ*. 2014;14(1):169.
35. Yingling S, Park YS, Curry RH, Monson V, Girotti J. Beyond cognitive measures: Empirical evidence supporting holistic medical school admissions practices and professional identity formation. *MedEdPublish*. 2018;7.
36. Cousans F, Patterson F, Edwards H, Walker K, McLachlan JC, Good D. Evaluating the complementary roles of an SJT and academic assessment for entry into clinical practice. *Adv Health Sci Educ Theory Pract*. 2017;22(2):401–413.
37. Davison I, McManus C, Taylor C. *Evaluation of GP Specialty Selection*. London: UCL; 2016.
38. Gardner AK, Dunkin BJ. Evaluation of validity evidence for personality, emotional intelligence, and situational judgment tests to identify successful residents. *JAMA Surg*. 2018;153(5):409–416.
39. Gardner A, Dunkin B. Making progress on identifying those who aren't making progress: using situational judgment tests to predict those at risk for remediation and attrition. *MedEdPublish*. 2018;7.
40. Libbrecht N, Lievens F, Carette B, Côté S. Emotional intelligence predicts success in medical school. *Emotion*. 2014;14(1):64–73.
41. Lievens F, Buyse T, Sackett PR. The operational validity of a video-based situational judgment test for medical college admissions: illustrating the importance of matching predictor and criterion construct domains. *J Appl Psychol*. 2005;90(3):442–452.
42. Lievens F, Sackett PR. The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. *J Appl Psychol*. 2012;97(2):460–468.
43. Lievens F. Adjusting medical school admission: assessing interpersonal skills using situational judgement tests. *Med Educ*. 2013;47(2):182–189.
44. Pashayan N, Gray S, Duff C, et al. 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*. 2016;38(2):e194–e200.
45. Paton LW, Tiffin PA, Smith D, Dowell JS, Mwandigha LM. Predictors of fitness to practise declarations in UK medical undergraduates. *BMC Med Educ*. 2018;18(1):68.
46. Patterson F, Lopes S, Harding S, Vaux E, Berkin L, Black D. The predictive validity of a situational judgement test, a clinical problem solving test and the core medical training selection methods for performance in specialty training. *Clin Med*. 2017;17(1):13–17.
47. Patterson F, Cousans F, Edwards H, Rosselli A, Nicholson S, Wright B. The predictive validity of a text-based situational judgment test in undergraduate medical and dental school admissions. *Acad Med*. 2017;92(9):1250–1253.
48. Patterson F, Tiffin PA, Lopes S, Zibarras L. Unpacking the dark variance of differential attainment on examinations in overseas graduates. *Med Educ*. 2018;52(7):736–746.
49. Patterson F, Martin S. UKCATSJT: a study to explore validation methodology and early findings. 2014. Published by the UCAT Board. https://www.ucat.ac.uk/media/1184/ukcat-sjt-validation-study-1.pdf.
50. Slaughter JE, Christian MS, Podsakoff NP, Sinar EF, Lievens F. On the limitations of using situational judgment tests to measure interpersonal skills: the moderating influence of employee anger. *Pers Psychol*. 2014;67(4):847–885.
51. Smith DT, Tiffin PA. Evaluating the validity of the selection measures used for the UK's foundation medical training programme: a national cohort study. *BMJ Open*. 2018;8(7):e021918.
52. Lievens F, Patterson F. The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced level high-stakes selection. *J Appl Psychol*. 2011;96:927–940.
53. Patterson F, Lievens F, Kerrin M, Munro N, Irish B. The predictive validity of selection for entry into postgraduate training in general practice: evidence from three longitudinal studies. *Br J Gen Pract*. 2013;63(616):e734–e741.
54. Patterson F, Rowett E, Hale R, et al. The predictive validity of a situational judgement test and multiple-mini interview for entry into postgraduate training in Australia. *BMC Med Educ*. 2016;16:87.

55. Patterson F, Baron H, Carr V, Plint S, Lane P. Evaluation of three short-listing methodologies for selection into postgraduate training in general practice. *Med Educ*. 2009;43(1):50–57.
56. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21(11):1539-1558.
57. Campion MC, Ployhart RE, MacKenzie WI. The state of research on situational judgment tests: a content analysis and directions for future research. *Hum Perform*. 2014;27(4):283-310.
58. Sterne JAC, Sutton AJ, Ioannidis JPA, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*. 2011;343:d4002.
59. McManus I, Dewberry C, Nicholson S, Dowell J, Woolf K, Potts H. Construct-level predictive validity of educational attainment and intellectual aptitude tests in medical student selection: meta-regression of six UK longitudinal studies. *BMC Med*. 2013;11(1):243.
60. Tiffin PA, Mwandigha LM, Paton LW, et al. Predictive validity of the UKCAT for medical school undergraduate performance: a national prospective cohort study. *BMC Med*. 2016;14(1):140.
61. Donnon T, Paolucci E, Violato C. The predictive validity of the MCAT for medical school performance medical board licensing examinations: a meta-analysis of the published research. *Acad Med*. 2007;82:100-106.
62. Emery JL, Bell JF. The predictive validity of the BioMedical Admissions Test for pre-clinical examination performance. *Med Educ*. 2009;43(6):557-564.
63. Christian MS, Edwards BD, Bradley JC. Situational judgment tests: constructs assessed and a meta-analysis of their criterion-related validities. *Pers Psychol*. 2010;63(1):83-117.
64. McDaniel MA, Morgeson FP, Finnegan EB, Campion MA, Braverman EP. Use of situational judgment tests to predict job performance: a clarification of the literature. *J Appl Psychol*. 2001;86(4):730-740.
65. Schwibbe A, Lackamp J, Knorr M, Hissbach J, Kadmon M, Hampe W. Medizinstudierendenauswahl in Deutschland [Selection of medical students in Germany]. *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz [Federal Health Gazette Health Research Health Protection]*. 2018;61(2):178-186.
66. Patterson F, Knight A, Dowell J, Nicholson S, Cousans F, Cleland J. How effective are selection methods in medical education? A systematic review. *Med Educ*. 2016;50(1):36-60.
67. Eva KW, Reiter HI, Trinh K, Wasi P, Rosenfeld J, Norman GR. Predictive validity of the multiple mini-interview for selecting medical trainees. *Med Educ*. 2009;43(8):767-775.
68. Tiffin PA, Carter M. *Understanding the Measurement Model of the UKCAT Situational Judgment Test: Summary Report*. London, UK: UKCAT Board; 2019:1-9. www.ucat.ac.uk/media/1183/understanding-the-measurement-model-of-the-ukcat-sjt.pdf. Accessed February 20, 2020.
69. Tiffin PA. *Understanding the Dimensionality and Reliability of the Cognitive Scales of the UK Clinical Aptitude test (UKCAT): Summary Version of the Report*. Durham, UK: Durham University; 2013:1-7.
70. Najim M, Rabee R, Sherwani Y, et al. The situational judgement test: a student's worst nightmare. *Adv Med Educ Pract*. 2015;6:577-578.
71. Singagireson S, Ramjeeawon N, Ravindra S, Shah N, Singh B. Is it fair for a junior doctor's deanery to be largely based on one test: a student's perspective. *Adv Med Educ Pract*. 2015;6:499-500.
72. UK Foundation Programme. UK Foundation Programme. UKFP; 2020. https://foundationprogramme.nhs.uk/. Accessed February 20, 2020.
73. Petty-Saphon K, Walker KA, Patterson F, Ashworth V, Edwards H. Situational judgment tests reliably measure professional attributes important for clinical practice. *Adv Med Educ Pract*. 2016;8:21-23.
74. Parrigon S, Woo SE, Tay L, Wang T. CAPTION-ing the situation: A lexically-derived taxonomy of psychological situation characteristics. *J Pers Soc Psychol*. 2017;112(4):642-681.
75. O'Connor PJ, Hill A, Kaya M, Martin B. The measurement of emotional intelligence: a critical review of the literature and recommendations for researchers and practitioners. *Front Psychol*. 2019;10:1116-1116.
76. Levashina J, Morgeson FP, Campion MA. They don't do it often, but they do it well: exploring the relationship between applicant mental abilities and faking. *Int J Select Assess*. 2009;17(3):271-281.
77. De Corte W, Sackett P, Lievens F. Designing pareto-optimal selection systems: formalizing the decisions required for selection system development. *J Appl Psychol*. 2011;96:(5):907-926.
78. Lievens F, Patterson F, Corstjens J, Martin S, Nicholson S. Widening access in selection using situational judgement tests: evidence from the UKCAT. *Med Educ*. 2016;50(6):624-636.
79. Juster FR, Baum RC, Zou C, et al. Addressing the diversity-validity dilemma using situational judgment tests. *Acad Med*. 2019;94(8):1197-1203.
80. Sam AH, Westacott R, Gurnell M, Wilson R, Meeran K, Brown C. Comparing single-best-answer and very-short-answer questions for the assessment of applied medical knowledge in 20 UK medical schools: Cross-sectional study. *BMJ Open*. 2019;9(9):e032550.
81. Guenole NR, Weekley JA, Ro S. Oral presentation: Scoring Keys and Measurement Models Not Required? SJT Predictions of Job Performance by Recursive Partitioning and its Variations. Oral Presentation. Paper presented at: 31st Annual Conference of the Society for Industrial and Organizational Psychology 2016; Anaheim, CA.
82. Aguinas H, Henle CA, Beaty JC Jr. Virtual reality technology: a new tool for personnel selection. *Int J Select Assess*. 2001;9:70-83.
83. Georgiou K, Gouras A, Nikolaou I. Gamification in employee selection: the development of a gamified assessment. *Int J Select Assess*. 2019;27(2):91-103.
84. Dowell J, Cleland J, Fitzpatrick S, et al. The UK medical education database (UKMED) what is it? Why and how might you use it? *BMC Med Educ*. 2018;18(1):6.
85. Lievens F, Corstjens J, Sorrel M, Abad F, Díaz J, Ponsoda V. The cross-cultural transportability of situational judgment tests: How does a US-based integrity situational judgment test fare in Spain? *Int J Select Assess*. 2015;23:(4):361-372.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Webster ES, Paton LW, Crampton PES, Tiffin PA. Situational judgement test validity for selection: A systematic review and meta-analysis. *Med Educ*. 2020;00:1-15. https://doi.org/10.1111/medu.14201