Adv in Health Sci Educ (2017) 22:401–413
DOI 10.1007/s10459-017-9755-4

Evaluating the complementary roles of an SJT and academic assessment for entry into clinical practice

Fran Cousans[1,2] · Fiona Patterson[1,3] · Helena Edwards[1] · Kim Walker[4,5] · John C. McLachlan[6] · David Good[7]

Received: 27 June 2016 / Accepted: 12 January 2017 / Published online: 8 February 2017

Abstract Although there is extensive evidence confirming the predictive validity of situational judgement tests (SJTs) in medical education, there remains a shortage of evidence for their predictive validity for performance of postgraduate trainees in their first role in clinical practice. Moreover, to date few researchers have empirically examined the complementary roles of academic and non-academic selection methods in predicting in-role

✉ Fran Cousans
fcc9@le.ac.uk

Fiona Patterson
f.patterson@workpsychologygroup.com

Helena Edwards
h.edwards@workpsychologygroup.com

Kim Walker
kim.walker@nes.scot.nhs.uk

John C. McLachlan
j.c.mclachlan@durham.ac.uk

David Good
dg25@cam.ac.uk

[1] Work Psychology Group, 27 Brunel Parkway, Pride Park, Derby DE24 8HR, UK

[2] Occupational Psychology, Department of Neuroscience, Psychology and Behaviour, University of Leicester, Leicester LE1 7RH, UK

[3] Department of Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, UK

[4] Scotland Foundation School Director, NHS Education for Scotland, Forest Grove House, Foresterhill Road, Aberdeen AB25 2ZP, Scotland, UK

[5] UK Foundation Programme Office, St Chad's Court, 213 Hagley Road, Edgbaston, Birmingham B16 9RG, UK

[6] Durham University, Holliday Building, Thornaby TS17 6BH, UK

[7] Department of Psychology, University of Cambridge, Kings College, Downing Street, Cambridge CB2 3EB, UK

performance. This is an important area of enquiry as despite it being common practice to use both types of methods within a selection system, there is currently no evidence that this approach translates into increased predictive validity of the selection system as a whole, over that achieved by the use of a single selection method. In this preliminary study, the majority of the range of scores achieved by successful applicants to the UK Foundation Programme provided a unique opportunity to address both of these areas of enquiry. Sampling targeted high (>80th percentile) and low (<20th percentile) scorers on the SJT. Supervisors rated 391 trainees' in-role performance, and incidence of remedial action was collected. SJT and academic performance scores correlated with supervisor ratings ($r = .31$ and .28, respectively). The relationship was stronger between the SJT and in-role performance for the low scoring group ($r = .33$, high scoring group $r = .11$), and between academic performance and in-role performance for the high scoring group ($r = .29$, low scoring group $r = .11$). Trainees with low SJT scores were almost five times more likely to receive remedial action. Results indicate that an SJT for entry into trainee physicians' first role in clinical practice has good predictive validity of supervisor-rated performance and incidence of remedial action. In addition, an SJT and a measure of academic performance appeared to be complementary to each other. These initial findings suggest that SJTs may be more predictive at the lower end of a scoring distribution, and academic attainment more predictive at the higher end.

Keywords Situational judgement tests · Academic attainment · Predictive validity · Trainee physicians · Supervisor ratings · In-role performance

Introduction

Historically, medical selection has been based on academic attainment (Ferguson et al. 2002), and a wealth of evidence offers consensus that this is an effective predictor of performance during medical education and training (Ferguson et al. 2014; Puddey and Mercer 2014). However, current research shows that recruiting physicians solely on the basis of academic attainment is likely to neglect important non-academic attributes required for success during clinical practice (Patterson and Ferguson 2010; Patterson et al. 2015a). In addition, considering postgraduate contexts, applicants for trainee physician roles are relatively homogeneous (i.e. high performing) academically, which can make differentiating between applicants on the basis of academic achievement challenging, and potentially inaccurate (McManus et al. 2008). Conceptually, therefore, it appears necessary for non-academic attributes, in addition to academic attainment, to be assessed throughout physicians' medical career progression (Patterson et al. 2015b). For this reason, in both undergraduate and postgraduate medical settings internationally, multi-method approaches to selection are increasingly used. These typically combine methods which assess academic attainment and non-academic attributes. However, to date there remains a relative dearth of empirical evidence which has assessed the value of combining such methods *in practice*, as few researchers have examined the complementary roles of different selection methods in predicting in-role performance (Prideaux et al. 2011; Patterson et al. 2016).

Predictive validity of SJTs in postgraduate medical training

Recently, situational judgement tests (SJTs) have emerged as reliable measures of non-academic attributes in medical settings (Patterson et al. 2015a). An SJT tests individuals' judgements about responses to professional dilemmas which they may encounter in a target role. Internationally, extensive literature demonstrates the reliability, validity and stakeholder acceptability of SJTs across a range of occupations, including in the context of medical selection (Hänsel et al. 2010; Patterson et al. 2013; Patterson 2013). However, although construct validity and reliability evidence of SJTs exists at postgraduate level for some medical specialties in the UK including General Practice and Core Medical Training (Patterson et al. 2017; Lievens and Patterson 2011), there is currently no predictive validity research from the UK or elsewhere at the point of entry into medical graduates' first role in clinical practice.

The UK foundation programme

One postgraduate training programme which employs a multi-method approach to recruiting trainee physicians, including an SJT, is the UK foundation programme (UKFP). Annually, the UKFP appoints approximately 8000 medical graduates to their first role as practising trainee physicians. A unique feature of the programme's process of assigning trainees to positions is that all applicants are *ranked* on the basis of their combined performance on an SJT and an educational (academic) performance measure (EPM). Importantly, the programme's approach to assigning training places is based on *matching and allocation*; such that applicants with the highest ranking are most likely to receive their first choice of training post, and in theory, all applicants can be appointed a post.

A major limitation in selection research is that outcome data are often unavailable for low scoring applicants since these individuals are less likely to be offered a position. This creates restriction of range in any analysis of the predictive validity of selection methods. This is problematic as it is therefore not possible to draw conclusions about how well the method predicts the performance of individuals at the bottom end of the distribution (Sackett and Ostgaard 1994; Sackett et al. 2007). However, the UKFP's approach to place allocation offers a unique opportunity to assess the predictive validity of the academic and non-academic selection methods, using an almost complete range of applicant scores. Only a very small number of applicants who score at the extreme lowest end of the population's distribution *and* who do not succeed in a subsequent face-to-face review do not receive a training position on the UKFP (fewer than 5%) (UKFPO 2015). As such, successful applicants' scores on the SJT and EPM span the great majority of available scores, which can then be compared with in-role performance.

This study aimed to address the gaps in existing evidence regarding the predictive validity of SJTs in medical selection for performance of trainee physicians in their first role in clinical practice, and to evaluate the complementary roles of two methods (an SJT and a measure of academic performance) in a postgraduate selection system.

Research questions

1. What is the predictive validity of an SJT for trainee physicians' in-role performance?
2. To what extent does an SJT complement academic performance in predicting trainee physicians' in-role performance?

Method

Sample

Participants were postgraduate trainees from five of the 20 UK foundation schools (training institutions) who began their clinical placements in 2013, following the first 'live' year of the new recruitment system into the programme in 2012. The schools were selected to provide geographical representation across England, Wales and Scotland. The five specific schools were also selected in order to ensure inclusion of higher and lower overall selection scores. For practical, cost and administrative reasons it was not possible to obtain outcome performance data for the entire cohort (N = 8162). As there are relatively few poor performers on the SJT at application, we sought to oversample the low-scoring population so that their performance in practice could be analysed with a large enough sample, which would be unlikely to be obtained if a random selection of scorers were targeted. High scorers provided a case comparison. The inherent advantage in this approach to sampling in the context of SJT research is that it takes into consideration that the relationship of SJT scores with outcome criteria may be non-linear, as well as being suitable for exploratory research such as this (Preacher et al. 2005). A high or low score was defined as greater than the 80th percentile, or lower than the 20th percentile, respectively.

From the population of trainees who had applied to the five foundation schools for the 2013 UKFP, 938 were identified as having suitable SJT scores to be included in the sample. Ethical approval was sought and trainees consented approval for anonymous data to be reviewed for research purposes during their application to the programme. A unique ID code was used to match trainee physicians' questionnaires to SJT scores and demographic data.

Predictor measures

Situational judgement test

The SJT was developed in line with best practice (Lievens et al. 2008), using a detailed analysis of the role of a trainee physician and review with subject matter experts (Patterson et al. 2010). The SJT was implemented into operational recruitment in 2012, following piloting that demonstrated its reliability for use in this context (Patterson et al. 2011). Participants sat the paper-and-pencil SJT in invigilated conditions at their medical schools on specified administration dates and SJT data were provided by the UK Foundation Programme Office. The SJT demonstrates sufficient item- and test-level results, with mean reliability coefficients across test versions ranging from $\alpha = .69$ to .72 (Patterson et al. 2014, 2015c).

Participants completed one of three versions of the SJT between December 2012 and January 2013. Each test paper consisted of 60 operational items. Test versions were statistically equated for difficulty using a chained linear equating process (Kolen and Brennan 2014), to ensure that candidates' scores were comparable across paper versions. The equated SJT scores were transformed into points on a 0-50 scale using a linear transformation.[1]

[1] Note that although theoretically it is possible for the SJT score range to be 0–50, the observed score range is much closer to that for the EPM (34–50).

Academic (educational) performance measure

The EPM was calculated based on a combined score for knowledge and skills performance over the first four years of candidates' undergraduate degree, the range of which was between 34 and 43 points. The EPM gives additional points for further degrees (up to five points), and publications, presentations, and prizes (maximum of two points). The total available range of EPM scores was therefore 34–50. EPM data were provided by the UK Foundation Programme Office.

Outcome measures

Supervisor ratings

In-role performance data were gathered towards the end of the training year in summer 2014, so that supervisors could report on trainee physicians' performance throughout the course of the year. In line with best practice (Lievens et al. 2005), a bespoke questionnaire was designed which criterion-matched items to behavioural performance indicators of professional attributes measured by the SJT. The questionnaire consisted of 32 items in total, spanning the professional attributes (Commitment to Professionalism, Coping with Pressure, Problem Solving and Decision Making, Patient Focus, and Working Effectively as Part of a Team). Example items include "Was trustworthy, reliable and responsive" (Commitment to Professionalism) and "Took time to build relationships with patients" (Patient Focus).

Supervisors rated trainees' performance on a Likert scale of 1 ('Needed Significant Development') to 6 ('Clear Area of Strength') and a mean of all 27 items was created ('Supervisors' overall score').

Supervisors' roles included Foundation Programme Directors, Clinical Supervisors and Educational Supervisors. The amount of time supervisors had supervised the trainee they were reporting on ranged from fewer than four months to over 12 months.

Cronbach's alpha shows high internal reliability of the questionnaire completed by supervisors ($\alpha = .94$). A principal components factor analysis of questionnaire scores showed that a single factor explained 69% in the low scoring group and 76% in the high scoring group. The majority of the trainees were rated using only one or two points on the rating scale across the entire questionnaire. Together this suggests that supervisors did not differentiate greatly between the different attributes, so comparisons between scores on individual attributes were unlikely to be meaningful. Therefore, supervisors' overall score was used as the single outcome variable during analyses.

Remedial action

Supervisors were asked to record whether participants had been subject to remedial action during the course of their training (a dichotomous variable). Remedial action is implemented for physicians performing poorly on both clinical and non-clinical skills. Remedial actions include one-to-one training, additional learning, simulation and coaching (Cleland et al. 2013).

Results

Descriptive statistics

Questionnaires were returned from 447 trainees (47.7% response rate). Fifty-six cases were removed due to unmatchable ID codes and/or less than 50% of the survey being completed, resulting in a total of 391 questionnaires suitable for analysis. Sample demographics and demographics for the entire 2013 population (for comparison) are presented in Table 1.

The mean age of both the trainee sample and the 2013 cohort as a whole was 26. Table 1 indicates that demographically, the study's sample is similar to the entire applicant cohort. This confirms that the sampling method enabled the identification and examination of predictor and outcome variables for high and low scorers, without artificially increasing or decreasing any demographic indicators within the sample.

Descriptive statistics for the predictor and outcome measures are displayed in Table 2.

In the sample of matched trainee physicians, the SJT scores intentionally reflect a bimodal distribution, whereas the EPM scores span the full range of available scores. During the application process, applicants' SJT and EPM scores are combined to create a total application score. Generally those in the high scoring SJT group received higher total application scores and those in the low scoring SJT group received lower total application scores, although this was not universally the case. As such, the total application score distributions span nearly the full range of available total application scores, despite the exclusion of the mid-range SJT scores. Figure 1 shows the distribution of total scores for the high and low scoring SJT groups.

Predictive validity of SJT and EPM

Given the non-normal distributions within the sample, non-parametric analyses were conducted. There was a significant, positive correlation between the SJT and total EPM scores ($r_s = .46$, $p < .01$). The direction and magnitude of this correlation changes when

Table 1 Trainee validity sample and 2013 applicant population demographics

	Total 2013 applicant sample		High scoring SJT group		Low scoring SJT group		Total sample	
	N	%	N	%	N	%	N	%
Male	3515	43.1	45	28.1	124	53.7	169	43.2
Female	4555	55.8	113	70.6	106	45.9	219	56.0
Did not disclose gender	92	1.1	2	1.3	1	0.4	3	0.8
Asian	1556	19.1	13	8.1	61	26.4	74	18.9
Black	241	3.0	0	.0	17	7.4	17	4.3
Chinese	364	4.5	2	1.3	18	7.8	20	5.1
Mixed	313	3.8	5	3.1	8	3.5	13	3.3
Other	264	3.2	2	1.3	11	4.8	13	3.3
White	5180	63.5	133	83.1	110	47.6	243	62.1
Did not disclose ethnicity	244	3.0	5	3.1	6	2.6	11	2.8

Table 2 Descriptive statistics for predictor and outcome measures

	High scoring SJT group					Low scoring SJT group				
	N	Min	Max	Mean	SD	N	Min	Max	Mean	SD
Predictor variables										
SJT score	160	43.40	48.70	45.23	1.26	231	26.10	37.90	34.88	2.41
EPM score	160	34.00	50.00	42.58	4.09	231	34.00	47.00	38.19	3.01
Supervisors' overall score	160	2.70	6.00	5.05	0.73	231	1.15	6.00	4.64	0.95
Remedial action	2	1.3	14	6.1						
No remedial action	155	98.7	215	93.9						

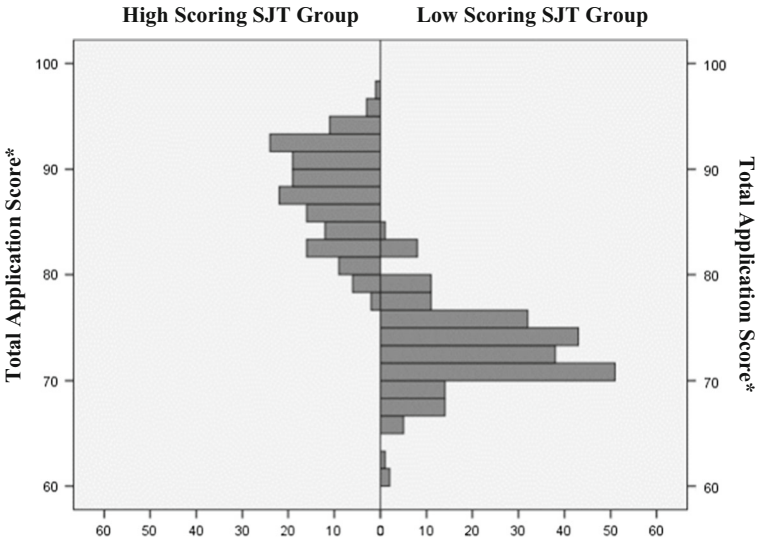

Fig. 1 Distribution of total application scores in the trainee sample

broken down by high or low scoring SJT group. In the low scoring group, the relationship between SJT and total EPM scores was $r_s = -.02$, $p > .05$. In the high scoring group however, $r_s = .20$, $p < .05$.

Supervisor ratings

An independent samples Mann–Whitney U test revealed significant differences in mean SJT scores between the two sample groups, with trainees with higher SJT scores receiving significantly higher supervisor ratings than those with lower SJT scores ($U = 13{,}806.50$, $Z = -4.26$ $p < .001$, $r = .22$). Spearman's correlation coefficients for SJT and EPM scores with supervisor ratings are reported in Table 3, showing that while both EPM and SJT scores correlate with supervisor ratings of performance for the sample as a whole, the relationship is only significant between EPM and supervisor ratings of performance for the

Table 3 Spearman's correlation coefficients for SJT and EPM scores with supervisor ratings

	N	EPM	SJT
High scoring SJT group	160	.29*	.11
Low scoring SJT group	231	.11	.33*
Whole Sample	391	.28*	.31*

* $p < .01$

Table 4 Mann Whitney *U* test to assess remedial action

	U	*Z*	Effect size (r)
EPM	1921*	−2.39	.12
SJT	1358**	−3.67	.19
Total application score (SJT + EPM)	1395*	−3.58	.18

* $p < .05$, ** $p < .01$

high scoring group, and between the SJT and supervisor ratings of performance for the low scoring group.

Remedial action

Mann–Whitney *U* tests were conducted to compare the SJT and EPM score distributions for trainees that did and did not receive remedial action during the course of the foundation year. Those that received remedial action scored significantly lower on the EPM, the SJT, and the total application score (see Table 4).

A χ^2 test for independence with Yates Continuity Correction indicated a significant association between presence of remedial action and score group (high or low SJT scores), [X^2 (1, n = 386) = 4.34, $p = .04$, phi = .12]. Whilst instances of remedial action were rare (4.1% in the sample as a whole), trainees who had received low SJT scores were almost five times more likely to receive remedial action than those receiving high SJT scores (6.1 and 1.3% respectively).

Discussion

Although there is extensive evidence confirming the predictive validity of SJTs in medical contexts (Lievens et al. 2012; Lievens 2013; Patterson et al. 2008, 2013, 2015a), there remains a relative shortage of evidence for the predictive validity of SJTs for performance for postgraduate trainees in their first role in clinical practice. Moreover, it is common practice in postgraduate medical settings internationally to combine measures of academic attainment and non-academic attributes in selection (Patterson et al. 2015b, 2016); however to date few researchers have empirically examined the complementary nature of these different selection methods in predicting in-role performance in practice (Prideaux et al. 2011; Patterson et al. 2016). The almost full range of scores achieved by successful applicants in this study provided a unique opportunity to address both of these areas of

enquiry. The results provide promising evidence to support the predictive validity and complementary contribution of an SJT in addition to indicators of academic attainment associated with in-role performance.

Predictive validity of the situational judgement test

Higher SJT scores were associated with higher supervisor ratings of trainee physicians' performance, and had approximately five times lower incidence of remedial action than the low scoring group. These early findings provide encouraging evidence for the validity of an SJT for recruitment into trainee physicians' first role in clinical practice, for predicting performance on non-academic criteria. The correlation coefficients in this study are comparable to other predictive validity studies of SJTs, as identified by meta-analyses (McDaniel et al. 2001, 2007).

Complementary roles of an SJT and academic performance

These preliminary results suggest that the combination of an SJT and a measure of academic attainment may enhance the predictive validity of selection system across the full range of applicant scores, showing the complementary roles of both methods in a postgraduate selection system. The present study indicates that both the SJT and EPM may have a non-linear relationship with supervisor ratings of in-role performance. The SJT only correlated significantly with supervisor ratings of performance at the lower end of SJT scores, whereas performance on the EPM correlated with supervisor ratings of performance only in the higher scoring SJT group. It is notable that the effect size of the SJT and EPM's relationships with the outcome criteria is approximately equal in their appropriate range ($r = .29$ for the SJT, $r = .33$ for the EPM), which provides support for the use of multiple methods in a selection system.

Practically, these findings imply that the SJT may be best used to identify candidates who are more likely to struggle in clinical practice. By contrast, the EPM appears to be best associated with non-academic performance during clinical practice at the highest end of the score distribution (i.e. those that perform *very* well on the EPM seem to be stronger in terms of their non-academic performance than those who perform quite well on the EPM). Indeed, the way that scores are allocated in the EPM means that applicants who get the highest marks are those that have strived to gain extra credit through publications, additional degrees, presentations and prizes; rather than those who are simply the most gifted academically. This proposition is supported by Patterson et al.'s (2015c) work.

The use of both academic and non-academic selection methods is therefore likely to be particularly beneficial in postgraduate medical recruitment (rather than selection) systems where there are frequently similar numbers of applicants to places available, as this combination of methods may allow for differentiation between applicants at both the high and low ends of the distribution. As such, this study provides evidence for the practical value of using multiple methods that target different selection criteria in a postgraduate medical recruitment system.

Implications for theory

Why might SJTs be more predictive at the lower end of the score distribution? Current theoretical developments in this area suggest that SJTs measure implicit trait policies

(ITPs) (Motowidlo et al. 2006), which may explain why SJTs are best placed to identify those likely to struggle during clinical practice (Patterson et al. 2015a). In the context of healthcare education and practice, prosocial ITPs are beliefs about the professional utility of acts which express compassion, caring, and respect for patients. For example, making a judgment that generally being agreeable (towards a patient, a colleague or a supervisor) may be a more successful strategy in dealing with a situation than being disagreeable. As such, SJTs may be able to identify applicants with ITPs fundamentally unsuited to working in a healthcare context, as arguably prosociality is a minimum requirement for any healthcare professional (see Patterson et al. 2015a for a discussion). In terms of implications for practice, recently researchers have suggested that SJTs may be best suited to 'selecting out' candidates, as an initial sifting tool to screen out those at the lower end of the distribution who do not have suitably prosocial ITPs to work in healthcare. Comparatively (and complementarily) measures of cognitive or academic ability may be most appropriately used to 'select in' at the top end of the distribution in the latter stages of a recruitment or selection system (Patterson et al. 2016).

Limitations and recommendations for future research

Sampling technique

Both a strength and a limitation of the current study is the sampling technique used to identify the bands of high and low scorers on the SJT. This method was beneficial because it allowed a direct comparison of the two applicant groups, and is appropriate for analysing non-linear relationships between predictor and outcome variables and exploratory analysis (Preacher et al. 2005). However, this approach inevitably excluded applicants with mid-range SJT scores from the analysis, so conclusions cannot be drawn from the current data about the performance of these individuals. This sampling approach prevented the assessment of the variance in in-role performance predicted by the SJT and EPM, and the incremental predictive validity of each tool over and above each other, which would be of value practically. Similarly, the bimodal distribution of the data prohibited the statistical analysis of the extent of the non-linear relationship between the predictor and outcome variables. Future research should aim to collect parametric data which allows for hierarchical regression and non-linear model fit analyses to be conducted.

Outcome measures

Remedial action may be implemented for trainees on the basis of a range of issues resulting in their poor performance being highlighted. These are often non-academic/clinical skills such as prioritisation, time management and communication. Data were not collected about the nature of the incidents which led to trainees receiving remedial action, and as a result it was not possible to assess how the SJT and EPM predicted remedial action as a result of non-academic/non-clinical or academic/clinic errors, respectively.

The questionnaire which collected supervisor ratings of performance focused on non-academic outcomes, criterion-matched to the SJT. As such, no 'purely' academic outcome measures were present in the study. This may have reduced the apparent predictive power of the EPM, as to assess the predictive validity of a selection tool in the most meaningful way, it should be criterion-matched with outcome measures (Lievens et al. 2005). Nonetheless, the EPM was more predictive of supervisor-rated performance at the highest

end of SJT scorers, which indicates that it does still predict non-academic performance during clinical practice.

Longitudinal follow up

Longitudinal studies are necessary to assess the predictive validity of the recruitment methods both during the two-year UKFP, and into later clinical practice. This study provides an important initial step in gathering such longitudinal data, and findings would be strengthened by subsequent follow-up of the same sample of trainees into specialty training and beyond, as well as by extending the present study with a wider, normally distributed, sample population.

Conclusions

The present study provides initial evidence that an SJT for entry into the role of post-graduate trainee has good predictive validity of supervisor-rated performance, as well as incidence of remedial action. Moreover, this study provides the first empirical evidence for the complementary roles of an SJT and a measure of academic attainment in recruiting trainee physicians into their first role in clinical practice. Our preliminary data suggest that a non-linear relationship exists between the two selection methods with performance during trainees' first year of clinical practice, such that the SJT has greater predictive validity for performance lower end of the distribution, and the EPM has greater predictive validity for those scoring at the higher end.

Acknowledgements The authors would like to thank Health Education England (HEE) who provided the funding to support this validation project as well as the UK Foundation Programme Office (UKFPO) and the UK Medical Schools Council (MSC) who supported the data access and collection. We are also grateful to the five Foundation Schools (Health Education North West Thames, North Western, Scotland, Wales and Yorkshire and the Humber), their Foundation School Managers and Supervisors who assisted with distributing and completing the questionnaires. Particular thanks also go to the steering group members for this project; Joanne Marvel (HEE), Sharon Witts (UKFPO), Katie Petty-Saphon (MSC) and Siobhan Fitzpatrick (MSC). The authors would also like to thank Lara Zibarras who reviewed the final iterations of this manuscript and provided helpful feedback.

Compliance with ethical standards

Conflict of interest FC, FP and HE are affiliated with Work Psychology Group which provides consulting advice and support to Health Education England on selection methodology. However, Work Psychology Group does not receive royalties for any methodology used.

References

Cleland, J., Leggett, H., Sandars, J., Costa, M. J., Patel, R., & Moffat, M. (2013). The remediation challenge: Theoretical and methodological insights from a systematic review. *Medical Education, 47*(3), 242–251.

Ferguson, E., James, D., & Madeley, L. (2002). Factors associated with success in medical school: Systematic review of the literature. *BMJ, 324*(April), 952–957. doi:10.1136/bmj.324.7343.952.

Ferguson, E., Semper, H., Yates, J., Fitzgerald, J. E., Skatova, A., & James, D. (2014). The "dark side" and "bright side" of personality: When too much conscientiousness and too little anxiety are detrimental with respect to the acquisition of medical knowledge and skill. *PLoS ONE*. doi:10.1371/journal.pone.0088606.

Hänsel, M., Klupp, S., Graupner, A., Dieter, P., & Koch, T. (2010). Dresden faculty selection procedure for medical students: What impact does it have, What is the outcome? *GMS Zeitschrift Für Medizinische Ausbildung*, *27*(2), Doc 25. doi:10.3205/zma000662.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking*. New York, NY: Springer New York. doi:10.1007/978-1-4939-0317-7.

Lievens, F. (2013). Adjusting medical school admission: Assessing interpersonal skills using situational judgement tests. *Medical Education, 47,* 182–189. doi:10.1111/medu.12089.

Lievens, F., & Patterson, F. (2011). 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. *The Journal of Applied Psychology, 96*(5), 927–940. doi:10.1037/a0023496.

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

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review, 37,* 426–441. doi:10.1108/00483480810877598.

Lievens, F., Buyse, T., Sackett, P. R., & Connelly, B. S. (2012). The effects of coaching on situational judgment tests in high-stakes selection. *International Journal of Selection and Assessment, 20*(3), 272–282. doi:10.1111/j.1468-2389.2012.00599.x.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Predicting job performance using situational judgment tests: A clarification of the literature. *Journal of Applied Psychology, 86*(4), 730–740.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60,* 63–91. doi:10.1111/j.1744-6570.2007.00065.x.

McManus, C., Woolf, K., & Dacre, J. E. (2008). Even one star at A level could be "too little, too late" for medical student selection. *BMC Medical Education, 8,* 16. doi:10.1186/1472-6920-8-16.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology*, *91*(4), 749.

Patterson, F. (2013). Selection into medical education, training and practice (pp. 383–397).

Patterson, F., & Ferguson, E. (2010). Selection for medical education and training. In *Understanding medical education: Evidence, theory and practice* (pp. 352–365). doi:10.1002/9781444320282.ch24.

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

Patterson, F., Archer, V., Kerrin, M., Carr, V., Faulkes, L., Coan, P., & Good, D. (2010). *FY1 job analysis report: Improving selection to the foundation programme*.

Patterson, F., Archer, V., Kerrin, M., Good, D., Carr, V., Faulkes, L., & Stoker, H. (2011). *Improving selection to the foundation programme. Appendix F. Report of the SJT pilots: Design and evaluation of a situational judgment test for selection to the foundation programme. Final report.*

Patterson, F., Ashworth, V., Murray, H., Empey, L., & Aitkenhead, A. (2013a). *Analysis of the situational judgement test for selection to the foundation programme 2013. Technical report.*

Patterson, F., Lievens, F., Kerrin, M., Munro, N., & Irish, B. (2013b). The predictive validity of selection for entry into postgraduate training in general practice: Evidence from three longitudinal studies. *British Journal of General Practice, 63*(616), 734–741. doi:10.3399/bjgp13X674413.

Patterson, F., Murry, H., Baron, H., Aitkenhead, A., & Flaxman, C. (2014). *Analysis of the situational judgement test for selection to the foundation programme 2014. Technical report.*

Patterson, F., Zibarras, L., & Ashworth, V. (2015a). Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100. *Medical Teacher, 00*(00), 1–15. doi:10.3109/0142159X.2015.1072619.

Patterson, F., Prescott-Clements, L., Zibarras, L., Edwards, H., Kerrin, M., & Cousans, F. (2015b). Recruiting for values in healthcare: a preliminary review of the evidence. *Advances in Health Sciences Education*. doi:10.1007/s10459-014-9579-4.

Patterson, F., Kerrin, M., Edwards, H., Ashworth, V., & Baron, H. (2015c). *Validation of the F1 selection tools*.

Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., & Cleland, J. (2016). How effective are selection methods in medical eduction and training? Evidence from a systematic review. *Medical Education*.

Patterson, F., Lopes, S., Harding, S., Vaux, E., Berkin, L., & Black, D. (2017). 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. *Clinical Medicine*, *17*(1), 13–17.

Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: a critical reexamination and new recommendations. *Psychological Methods, 10*(2), 178–192. doi:10.1037/1082-989X.10.2.178.

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

Puddey, I. B., & Mercer, A. (2014). Predicting academic outcomes in an Australian graduate entry medical programme. *BMC Medical Education*. doi:10.1186/1472-6920-14-31.

Sackett, P. R., & Ostgaard, D. J. (1994). Job-specific applicant pools and national norms for cognitive ability tests: Implications for range restriction corrections in validation research. *The Journal of Applied Psychology*, *79*(5), 680–6844. http://www.ncbi.nlm.nih.gov/pubmed/7989275.

Sackett, P. R., Lievens, F., Berry, C. M., & Landers, R. N. (2007). A cautionary note on the effects of range restriction on predictor intercorrelations. *Journal of Applied Psychology*. doi:10.1037/0021-9010.92.2.538.

UKFPO. (2015). *FP/AFP 2015 applicant's handbook*.