International Journal of Assessment Tools in Education: Vol. 4, Issue 2, (2017) pp. 96-114

IJATE

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE

INTERNATIONAL JOURNAL of ASSESSMENT TOOLS in EDUCATON

"Research Article"

Developing a Proof-of-Concept Selection Test for Entry into Primary Teacher Education Programs

Robert M. Klassen,[1] Tracy L. Durksen,[2] Lisa E. Kim,[1] Fiona Patterson,[3,4] Emma Rowett,[3] Jane Warwick,[4] Paul Warwick,[4] and Mary-Anne Wolpert[4]

[1]Department of Education, University of York, York YO10, 5DD, United Kingdom
[2]University of New South Wales, Australia
[3]Work Psychology Group, UK
[4]University of Cambridge, UK

Abstract

The purpose of this article is to report on the development of a proof-of-concept situational judgment test (SJT) to assist in the selection of candidates for primary teacher education (ITE) programs. Nine development steps involving practising teachers, teacher educators, and applicants to ITE programs were carried out to establish target attributes and to develop content for the test. The results from administering the test to 124 primary ITE candidates showed a near-normal distribution, high levels of reliability, and significant positive correlations with a range of concurrently administered interview scores. We conclude with a description of the necessary next steps needed to implement evidence-supported teacher education selection processes in a range of international settings.

Article Info

Received
26 September 2016

Revised
19 November 2016

Accepted
23 November 2016

Key words
teacher selection;
initial teacher education;
situational judgment tests;
teacher effectiveness;
recruitment;
teacher characteristics

1. INTRODUCTION

Identifying and selecting the most promising prospective teachers has been a continuing challenge in educational research and practice for nearly 100 years (e.g., Knight, 1922; Staiger & Kane, 2015). Any selection process is built on an evaluation of data to make predictions about future effectiveness. Selecting candidates for initial teacher education (ITE) programs presents selectors with questions about the kinds of data to evaluate: Which characteristics of candidates should be evaluated? How can these characteristics be evaluated in a way that is reliable, valid, and fair? Are these characteristics associated with success in teacher education and teaching practice? The conventional selection approach for ITE programs is to ask candidates for some combination of academic transcripts, personal statements, letters of reference, and to participate in individual interviews. However, there is little evidence supporting the use of many conventional ITE selection procedures (Casey & Childs, 2011),

[1] Corresponding Author Phone: +44 07914 701260 Email: robert.klassen@york.ac.uk

2148-7456 /© 2017 DOI: 10.21449/ijate.275772

and furthermore, some selection methods-including interviews and letters of reference-may be unreliable and systematically biased against certain groups of candidates (McDaniel, Whetzel, Schmidt, & Maurer, 1994). In this proof-of-concept study, we report the development and initial evaluation of an innovative selection tool for use in selecting candidates for primary ITE programs.

1.1. The case for improving selection procedures into initial teacher education

High-performing education systems tend to place importance on developing effective ITE selection processes (Barber & Mourshed, 2007; Sahlberg, 2014; Sclafani, 2015), with selection methods that include evaluation of candidates' academic and non-academic attributes[1]. Researchers and policy-makers in a range of settings have called for improvements in ITE selection in efforts to improve teacher quality (Heinz, 2013; Thomson et al., 2011; UK House of Commons, 2012). In any jurisdiction, selection is necessary for three reasons: a) to make decisions about 'selecting in' when the number of applicants outweighs the number of available places, b) to make decisions about 'selecting out' in order to identify those candidates who may be unsuitable, and c) to provide a profile of candidates' strengths and weaknesses for future development. At the foundation of selection research is the belief that individuals vary in personal attributes and experiences, and that these individual differences are related to future behaviors in training and professional contexts.

Although almost all novice teachers become more effective with experience and professional training (Hanushek & Rivkin, 2011), their effectiveness relative to their peers remains quite stable over time (Atteberry, Loeb, & Wyckoff, 2015). That is, novice teachers' relative effectiveness is heterogeneous and is predictive of their future relative effectiveness, especially for those who initially display the highest and lowest levels of relative effectiveness (Atteberry et al.). Furthermore, although many candidates entering ITE programs will show growth in non-academic attributes (e.g., professional commitment and motivation) during the duration of their program, some candidates will show persistently low levels of professional commitment and motivation (e.g., Klassen & Durksen, 2014; Watt, Richardson, & Wilkins, 2014). Watt et al. (2014) traced the professional commitment and motivation of students from the beginning to the end of their ITE programs, and found that a sizable group-28% of participants in their study-began the program with low levels of motivation for teaching and maintained that profile until the end of the program. Given the relative stability of teacher effectiveness and non-academic attributes, selection methods used by ITE programs should make the best possible predictions about the motivation and effectiveness trajectories of prospective teachers.

1.2. Current approaches for ITE selection

Uncovering the within-teacher factors that lead to teacher effectiveness is at the heart of the ITE selection process. Although attempts have been made to improve and systematise selection practices, there is a dearth of valid tools to help admissions committees make these important selection decisions in ITE programs (Mikitovics & Crehan, 2002). Selection into ITE programs typically involves evaluation of three factors: (1) academic attributes (such as

[1] The term 'academic' attributes (sometimes referred to as 'cognitive' attributes) refers to variables that reflect reasoning skills (such as the Scholastic Aptitude Test, SAT) or academic achievement (e.g., GPA or past performance in particular academic areas). The term 'non-academic attributes' (sometimes referred to as 'non-cognitive' attributes) refers to within-person variables, which might include beliefs, motives, personality traits, and dispositions (e.g., Patterson, Zibarras, & Ashworth, 2016).

subject area knowledge using evidence from university transcripts and sometimes through a written response to a journal article); (2) background experience (using evidence from personal statements and reference letters); and (3) *non-academic* attributes (such as personality, motives, and dispositions using evidence from interviews, personal statements, and occasionally, personality tests).

Figure 1 provides a model with examples of how these three factors are measured and how they are linked to performance for selection into ITE programs. Although teacher education programs vary in the kinds of assessments that they use for assessing candidates, we know very little about the reliability, validity, and perceived fairness of these procedures. What links disparate selection methods together is the common goal to identify candidates who show higher, rather than lower, levels of academic and non-academic attributes.

In the UK, a recent survey of 74 university-based (ITE) providers (Klassen & Dolan, 2015) found that all programs assessed academic attributes through evaluation of university academic transcripts, and that almost all assessed non-academic attributes through a combination of individual and group interviews (97%), and evaluation of behaviour in group activities (62%). In North America, specific selection methods for ITE programs vary widely, but selectors typically rely on some combination of candidates' previous academic achievement, individual and group interview performance, personal statements, letters of reference, and in some cases, government-mandated standardized tests (Casey & Childs, 2007). Selection into highly competitive Finnish ITE programs includes evaluation of academic attributes such as academic achievement, but also non-academic attributes including personality and interpersonal skills (Sahlberg, 2014). Similarly, selection into competitive Singaporean ITE programs includes an evaluation of academic attributes such as grades and national exams, but also evaluation of non-academic attributes including motivation, passion, values, and commitment to teaching (Sclafani, 2015). Almost all selection approaches have the same goal—to identify candidates with the highest potential for success during the program and in teaching practice—but there is little evidence for reliability, validity, and fairness of these selection methods internationally (Hobson, Ashby, McIntyre, & Malderez, 2010).

1.3. Situational judgment tests

In fields outside of education, there has been a keen interest in the use of situational judgment tests (SJTs) for employee selection, but also for selection into professional training programs, especially in medicine (e.g., Patterson, Zibarras, & Ashworth, 2016). SJTs are a measurement method designed to assess candidates' judgments of the benefits and costs of behaving in certain ways in response to challenging contextualised scenarios. In some ways, SJTs resemble a conventional face-to-face interview where a scenario might be presented orally to candidates with an open-ended response format (e.g., *Describe what you would do if….*). SJTs, however, differ from conventional interviews in that a larger sample of scenarios can be administered to applicants, the scoring key can be standardized, and the tests can be used to screen large numbers of applicants economically and efficiently. The format of SJTs can be in paper-and-pencil, computer-administered, or video-based. The development of SJT content is typically based on job analysis and through gathering 'critical incidents' from those already in the job (Patterson et al., 2016). Experienced professionals, or 'subject matter experts,' are used to generate response options (Lievens et al., 2008). Final scoring keys, which indicate more and less effective response options, are established through consensus with a panel of experts.

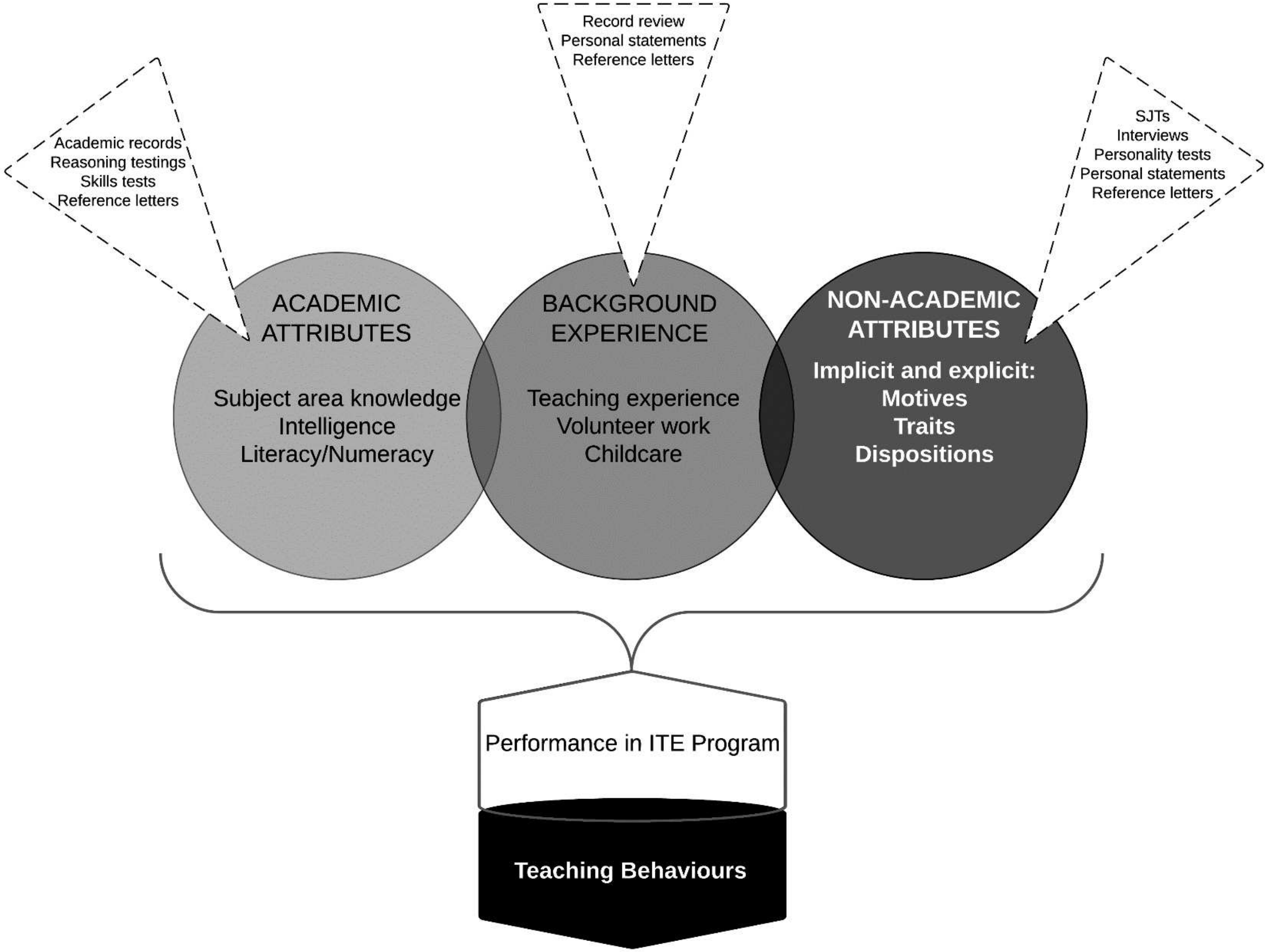

Figure 1. Model of relationship between academic attributes, background experience, and non-academic attributes in prediction of performance of ITE performance and teaching behaviors.

SJTs are designed to measure *implicit trait policies*; that is, the tendency individuals have to express traits in certain ways under particular contexts (Motowidlo & Beier, 2010). According to this theory-similarly conceptualised as tacit knowledge in Sternberg's theory of successful intelligence (e.g., Elliott, Stemler, Sternberg, Grigorenko, & Hoffman, 2011)-those who are more experienced in a particular job are more likely to implicitly understand optimal behavioral responses. However, novices with limited experience also have partial knowledge about effective response patterns, based on their implicit traits and understanding of the kinds of behaviors that are likely to be most appropriate in SJT scenarios (Motowidlo & Beier). In education, candidates for ITE programs have pre-existing beliefs about how to react to classroom challenges (e.g., how to manage classroom discipline issues), based on the procedural knowledge gained from their own life experiences, even when they do not have direct experience with teaching. These existing beliefs, or implicit trait policies, may change as candidates gain pedagogical knowledge and teaching experience, but remain as influences of teaching behaviors.

SJTs tend to display stronger face and content validity than conventional non-academic measures due to their close correspondence to the work-related situations that they describe (Whetzel & McDaniel, 2009).The interest in SJT methodologies is due to the promise of predictive validity (Patterson et al., 2016), with SJTs administered at admissions to medical school predicting job performance (r = .22) nine years later (Lievens & Sackett, 2012). In a recent meta-analysis on SJT validities and reliabilities, Christian et al. (2010) found SJTs measuring interpersonal attributes had a mean validity coefficient of .25, those measuring conscientiousness had a mean coefficient of .24, and heterogeneous composite SJTs showed a mean validity of .28. A previous large-scale meta-analysis of SJT validity (N = 24,756) using mostly concurrent validity studies showed a validity coefficient of .26 (McDaniel, Hartman, Whetzel, & Grubb, 2007).

Non-academic attributes can be measured using conventional, explicit measures of personality (e.g., 'How much is this statement like you?' *I am generally agreeable*) that are prone to socially desirable response patterns (Greenwald & Banaji, 1995; Johnson & Saboe, 2011). In contrast, SJTs can provide an indirect or implicit measure of what candidates view as appropriate ways of behaving in certain contexts (Motowidlo & Beier, 2010). Moreover, SJTs constructed in collaboration with expert practitioners are less susceptible to coaching effects and faking than many other kinds of selection tests because they are cognitively complex and are designed to measure implicit traits (Whetzel & McDaniel, 2009).

Researchers have also noted weaknesses in the research underpinning the development and use of SJTs for selection (e.g., Lievens, Peeters, & Schollaert, 2008). The vast majority of SJT validation studies have used a concurrent design with few studies establishing predictive validity (Campion, Ployhart, & MacKenzie, 2014). Although SJTs are often constructed to target particular attributes (e.g., professional integrity in medical selection; Patterson et al., 2016), their hypothesized factor structure is frequently not replicable in factor analysis (Lievens et al., 2008). In addition, internal consistency may be below conventional standards, and some SJTs have been shown to be prone to faking and coaching (Whetzel & McDaniel, 2009). SJTs are typically developed to reflect multiple dimensions, but because the content of individual items (scenarios) may reflect multiple dimensions, establishing the factor structure can be a challenge (Schmitt & Chan, 2006).

SJTs have been shown to predict performance in dentistry and medical training programs over and above cognitive measures (Lievens & Sackett, 2012; Patterson, et al, 2012). In the United States, SJTs were found to be a better predictor of lawyer effectiveness than the conventional tests used for selection into highly competitive law schools, and to be less prone to inter-group bias (i.e., race, gender) than other measures (Shultz & Zedeck, 2012). Overall, SJTs have shown strong concurrent validity, some evidence of predictive validity (Lievens & Patterson, 2011), and a higher degree of fairness (i.e., less systematic bias) than other selection methods (Shultz & Zedeck, 2012).

Current Study. SJTs are often designed deductively (top-down) to capture personality traits, but can also be designed to measure inductively-developed, contextualised non-academic attributes related to professional effectiveness. The current study describes the development and initial validation of a proof-of-concept SJT designed to be used for selection into primary level teacher education programs in the UK. Four research questions were posed:

(RQ1) Can a set of robust target attributes be established based on an inductive (bottom-up) approach?
(RQ2) Can an SJT developed for entry into primary ITE show acceptable psychometric properties?
(RQ3) Is the SJT a valid selection method (i.e., does the SJT show concurrent criterion-related validity with scores from the existing selection process)?
(RQ4) Do candidates view the SJT as fair and as a feasible selection method (i.e., does the test show face validity)?

2. METHOD AND RESULTS

The ITE selection SJT was designed to assess non-academic attributes required for success as a novice teacher in UK primary schools. We followed best-practice approaches to SJT development from the organizational psychology literature (Campion et al., 2014), and in particular, the approach used by Patterson et al., 2015 as part of their creation of selection tests used for medical training. Figure 2 illustrates the three phases and nine steps of the development process. In Phase 1, we developed the target attributes on which the content (scenarios and responses) of the SJT were based. We used an inductive approach with data gathered through observation of practising teachers, individual and focus group interviews with teachers and teacher educators, and questionnaires with teachers and teacher educators. An inductive approach to SJT development has been widely used in organizational psychology (Campion et al., 2014) and for developing selection tools for medical education (Patterson et al., 2016). In Phase 2, we created scenarios and responses for the SJT. In Phase 3, we carried out an initial validation of the SJT using concurrent data from current selection processes with participants from three ITE programs in the UK.

Steps 1-3: Identifying target attributes. Three steps were carried out to establish the target attributes for the SJT[1]. Defining the target attributes is an important step in developing SJTs, since creation of SJT content (scenarios and response options) is grounded in the target attributes. Step 1 consisted of full-day observations and in-depth interviews with two practising teachers in two schools. Step 1 was designed to provide an initial awareness of the activities and behaviors of the target teachers, inside and outside of the classroom. One teacher was a mid-career teacher and one was a newly-qualified teacher in her first year of practice after completing a teacher education program. A detailed summary report was produced describing the teachers' routines from the start of the day (e.g., 'up at 5 a.m., drive to gym') to the close of the day (e.g., 'as soon as child in bed, marking for 1 hour'). The purpose of Step 1 was not to provide an exhaustive or representative exploration of school life, but to

[1] Steps 1-3 were carried out for the development of an earlier version (for primary and secondary ITE applicants) of the SJT (see Klassen, Durksen, Rowett, & Patterson, 2014). In Step 4 we revised the target attributes created in Steps 1-3.

(re)familiarise the research team with the daily activities of teachers and the general functioning of schools.

In Steps 2 and 3, three focus group interviews were conducted in two schools (n = 18) and one university teacher education program (n = 10), and included practising teachers, school leaders, and teacher educators. Step 2 was designed to inductively identify the target attributes needed for successful novice teaching. The 28 expert participants were recommended by teacher education leaders and recruited from the pool of teachers and teacher educators who were involved in pre-service teacher supervision. We generated discussion using a critical incident approach where participants were encouraged to consider 'critical incidents' that led to positive or negative outcomes, e.g., *Think of a event where a newly-qualified teacher showed good (bad) judgment.* In addition, focus group participants were asked to generate and rate academic and non-academic attributes necessary for success for new teachers. Focus group data were collected and analysed using a content analysis approach. The focus group meetings resulted in the generation of 13 initial attributes (e.g., *caring, fairness, enthusiasm, reflection*) with behavioral descriptors.

Step 3 consisted of an iterative process of data reduction and integration led by three of the authors, and carried out through discussions with teachers and teacher educators about the importance of the 13 initial attributes (i.e., *How important are these attributes for new teachers?*). We used a multi-method consensus approach that integrated numerical ratings of the attributes with individual and group discussion of the relative importance of the attributes. In particular, we used a data reduction process that involved proposing clusters of domains to teacher and teacher educator focus groups and that asked *Which of these attributes are critical for the success in the teacher education program?* and *Which attributes are critical for the success of new teachers?* The 13 initial attributes were discussed individually and summarized into themes, or domains, with operational descriptors generated through discussion.

Phase 1: Establishing Target Attributes

After completion of the data reduction process, three composite domains-each consisting of two target attributes-emerged through further discussion and group consensus: *Empathy and Communication, Organisation and Planning, and Resilience and Adaptability*. The three composite domains were next evaluated for suitability to capture the key attributes specifically needed for novice teachers working in primary school contexts.

Step 4: Reviewing target attributes. Step 4 was conducted to evaluate and revise the target attributes specifically for the primary school environment. We posed three questions to seven experienced teacher educators from three UK university-based teacher education programs:

- *Do the three broad domains (and six target attributes) capture the non-academic attributes necessary for successful novice teaching at the primary school level?*
- *Are there any additional attributes that need considering?*
- *How do these attributes need adapting for a primary school teaching context?*

The review of target attributes resulted in retention of the three composite domains, but with a revision of the operational descriptors for a primary school environment. For example, the domain "Organisation and Planning" was broadened by consensus to include elements relating to managing competing priorities in order to capture the multiple demands primary school teachers face. Table 1 presents the three composite domains with the six target attributes and their descriptors. The domains generated in Steps 1-4 formed the foundation of the SJT content, and served as the basis for creating items (scenarios) and responses.

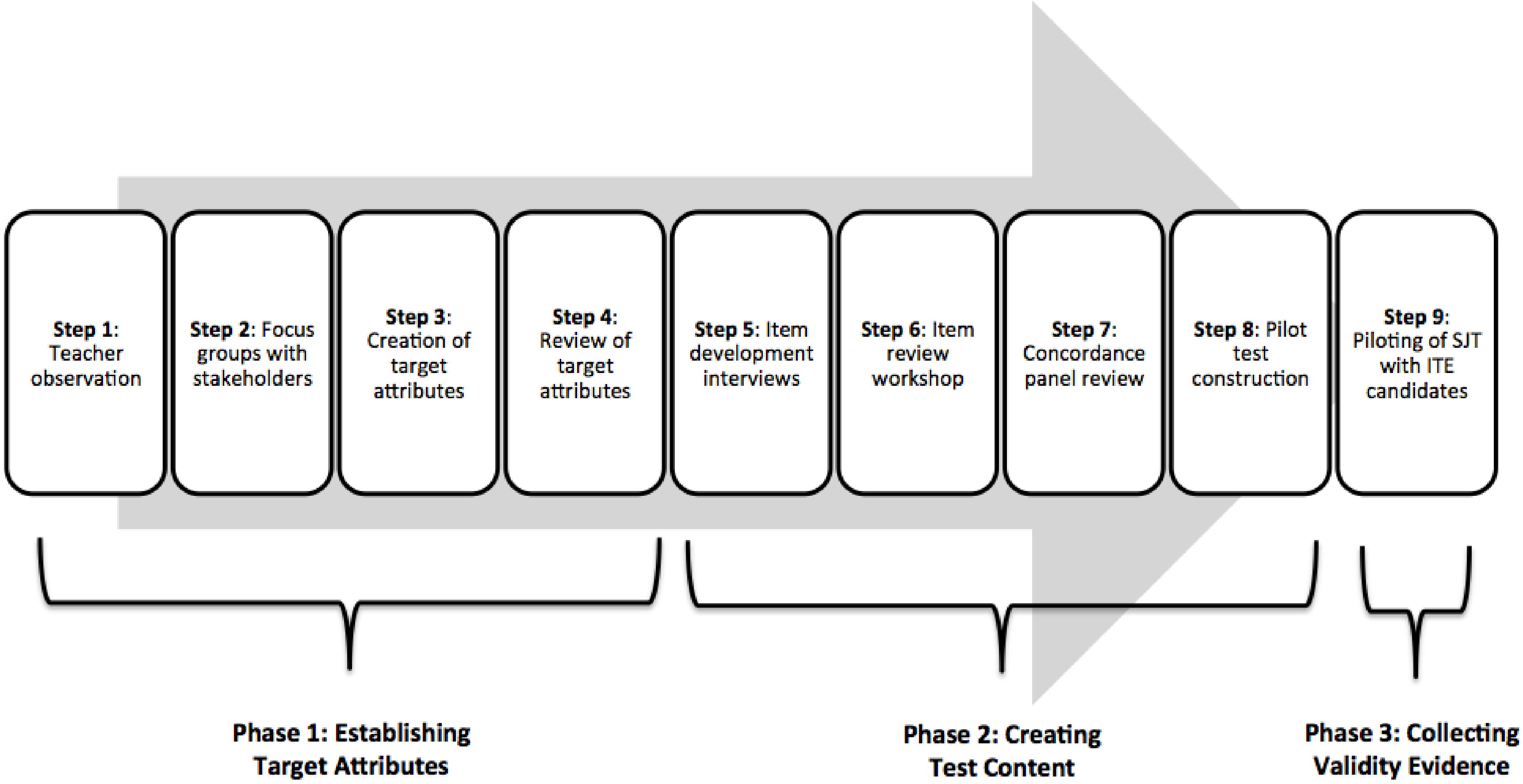
Step 1: Teacher observation
Step 2: Focus groups with stakeholders
Step 3: Creation of target attributes
Step 4: Review of target attributes
Step 5: Item development interviews
Step 6: Item review workshop
Step 7: Concordance panel review
Step 8: Pilot test construction
Step 9: Piloting of SJT with ITE candidates
Phase 1: Establishing Target Attributes
Phase 2: Creating Test Content
Phase 3: Collecting Validity Evidence

Figure 2. Nine steps of development of target attributes and pilot situational judgment test.

Table 1. Composite Domains and Target Attributes Identified for Teacher Selection SJT

Domain	Description
Empathy and Communication	Candidate demonstrates active listening, and engages in an open dialogue with both pupils and colleagues. Candidate seeks advice pro-actively and is responsive to both professional feedback and pupils' needs. Candidate has the ability to adapt the style of communication and nature of dialogue appropriately.
Organisation and Planning	Candidate has the ability to manage competing priorities and display time management and personal organisation skills effectively, using these skills to enhance positive learning interactions with pupils.
Resilience and Adaptability	Candidate demonstrates the capability to remain resilient under pressure. Demonstrates adaptability, and an ability to change lessons (and the sequence of lessons) accordingly where required. Candidate has an awareness of their own level of competence and the confidence to either seek assistance, or make decisions independently, as appropriate. Is comfortable with challenges to own knowledge and is not disabled by constructive, critical feedback. Uses effective coping strategies.

Phase 2: Creating Test Content

Phase 2 consisted of four steps (Steps 5 to 8) aimed at developing content for the SJT based on the target attributes.

Step 5: Item development interviews. Step 5 was conducted by trained interviewers (from an organizational behavior consulting firm) with practising teachers to develop scenarios and responses based on the identified target attributes. Eleven teachers who had experience working with novice teachers (i.e., as mentors of newly-qualified teachers) were individually interviewed in order to generate classroom scenarios and response options. A critical incidents method was used, whereby participants were asked to reflect on challenging situations that they had experienced as novice teachers or that they had observed when supervising novice teachers (Anderson & Wilson, 1997). Participants were guided to generate critical incidents related to the six target attributes. The resulting critical incidents were used as the basis for creating 54 SJT scenarios and responses. Table 2 presents an example SJT item that resulted from an item development interview.

Step 6: Item review workshop. A one-day workshop with eight experienced teachers from six UK primary schools (chosen for their involvement in supervising novice teachers), together with three teacher educators was held to review the 54 items (scenarios with associated response options) generated in Step 5. The workshop began with an introduction to item review principles and SJT attributes (e.g., *Is the item set in the correct context? Is the item set at an appropriate level for a novice teacher [not an experienced teacher]? Are the responses plausible? Does the content depend on specific knowledge* [which would unfairly discriminate against participants without a particular background]*?*). Participants were then arranged in pairs to review the 54 SJT items, followed by group work to revise problematic items. The workshop concluded with a calibration session where participants reviewed and discussed decisions made about content revision. The workshop resulted in an initial draft SJT consisting of all 54 items that were generated through item development interviews.

Step 7: Concordance panel review. In a concordance panel, test items are completed and evaluated by experts, and a scoring key is determined from a consensus of the experts (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006). A concordance panel review session was conducted to identify a level of scoring consensus between expert reviewers in order to conclude which items had the highest degree of scoring agreement and to establish a scoring key. The 11 participants in the concordance panel were 9 experienced teachers and 2 teacher educators who worked closely with trainee teachers in schools and teacher education programs. Panel members completed the SJT in a 2-hour session, and provided additional feedback on the suitability and relevance of the scenarios and response options. Based on the scoring consensus and feedback on the 54 items, 35 items were selected for piloting with ITE candidates.

Step 8: Pilot test construction. The items were further revised based on feedback from the concordance panel (Step 7) and piloted with its scoring key. The pilot version of the SJT consisted of 35 scenarios designed for ITE candidates to complete in one hour. Five items represented the Organisation and Planning composite domain, 12 items represented Empathy and Communication, and 18 items represented Resilience and Adaptability. In order to reduce potential coaching effects (e.g., Whetzel & McDaniel, 2009), we used two response formats: 22 items used a ranking format (i.e., *Rank responses to this situation in order of appropriateness*) using a 5-point scale, and 13 items used a multiple response format (e.g., *Choose the three most appropriate actions to take in this situation*). Test scoring used a near miss scoring approach: for ranking items, candidates received partial points for correct responses that were not in the optimal order. For example, four points were awarded to an item in correct position, three points for an item adjacent to correct position, two points for an item two positions away, and so on. For multiple response items, candidates received four points for each correct answer, giving a possible total of 12 points for each scenario.

Table 2. Example of SJT Scenario

You are teaching a lesson and have asked the students to individually complete an exercise that requires them to write down their responses. You have explained the exercise to the students and answered all of the questions that they have asked. As the students begin writing, one student, Ruby, starts to throw paper around and is clearly distracting the students sitting nearby. You know from previous incidents that Ruby often becomes frustrated when she does not understand how to complete activities, and that she often displays her frustration by being disruptive.

Choose the three most appropriate actions to take in this situation (**alternatively,** ***Rank the items in the most appropriate order)***

- Send Ruby out the class if she continues to be disruptive
- Ask Ruby if she understands what the activity requires her to do
- Check in five minutes to see if Ruby has made progress with the exercise
- Tell Ruby that you are disappointed in her behavior
- Ask Ruby's classmate to discreetly provide help
- Stop the exercise and discuss the classroom behavior plan with the whole class
- *etc*. (eight total response options)

Note. This is an example only, and is adapted from an item from the primary SJT.

Phase 3: Collecting Reliability and Validity Evidence

Step 9: Piloting of SJT with ITE candidates. The final step in the last phase of development consisted of piloting the SJT with participants at two UK university ITE programs during their interview day. Participants were volunteers who were asked during the interview day if they would be willing to spend one hour completing the SJT. Interview day administrators estimated that 60% of candidates volunteered to complete the SJT during the course of the interview day, which consisted of procedures such as group activities, a written task, and individual interviews. A total of 124 candidates agreed to complete the SJT. Most of the candidates were female (81%) and white British (97.5 %), with a mean age of 22.3 years (range 20-34 years).

Descriptive statistics. Analysis of the 35-item test scoring resulted in three items being dropped due to low item quality (low correlations with total test score), leaving 32 items for further analysis. The mean score of the test was 407.3 (*SD* = 33.19), with a range of 270 to 458. The difficulty level of the test was 76% (i.e., the mean score was 76% of the total possible score. As is conventional for SJTs, we did not calculate means, reliability coefficients, or validity coefficients for the individual domains (e.g., Lievens et al., 2008).

The reliability of the 32-item SJT (α = .79) compares favourably with other SJTs used in selection contexts (Whetzel & McDaniel, 2009). The mean test score was 407.3 (range 270 to 458) with a maximum possible score of 536. The distribution of the scores was near normal, with a slight negative skew, meaning that most candidates scored in the higher range of the test rather than the lower range.

Validity. We used interview scores for 108 participants provided by ITE program coordinators to test the SJT's concurrent validity. The seven scoring categories for the interview (scored on a 1-4 scale) were:

(1) ability to communicate in standard English
(2) pedagogical and subject knowledge
(3) reflections on experience
(4) understanding of education practice
(5) quality of thinking
(6) personal attributes and skills, and
(7) overall interview score.

Table 3 provides the means and standard deviations for the seven interview scores, and the correlations between the interview scores and total SJT score. The SJT showed significant positive correlations with each mean interview score ($.21 \leq r \leq .29$, $p < .01$), suggesting that the SJTs measured attributes that overlapped with the attributes measured by a wide range of interview indicators. The SJT showed a correlation of .29 with the overall interview score.

Candidate reactions. We also collected data on candidates' perceptions of fairness, feasibility, and reasonableness of using SJT as part of the selection process because candidates' perceptions of the selection process influence their opinions of the organisation (Walker et al., 2013). From a recruitment perspective, a teacher training program's ability to successfully recruit applicants is influenced by the perceptions of current and past applicants, who may share word-of-mouth accounts about the fairness of the selection process, ultimately influencing the success of recruiting the best possible candidates.

Candidates reported a range of test completion times, with 56% of candidates reporting a completion time of 40–60 minutes and 42% of candidates reporting a completion time of less than 40 minutes. Most candidates (79%) *agreed/strongly agreed* that the test was "clearly relevant for those applying for ITE", and 74% *agreed/strongly agreed* that the level of difficulty was appropriate for ITE candidates. A majority of candidates (76%) *agreed/strongly agreed* that the content of the SJT appeared to be fair. Given an opportunity for open-ended responses, candidates commented that the test was useful to "place themselves in real life situations" and "far more applicable to the type of teaching experienced in the classroom" compared to other selection tests that they had taken for admission into other ITE programs. A minority of candidates commented that the test was too long and that, in some scenarios, it was difficult to judge the appropriate responses in the absence of additional information.

Table 3. Correlations Between Interview Scores and SJT Total Score

	Interview domains						
	Ability to communicate	Pedagogical & subject knowledge	Reflections on experience	Understanding of education	Quality of thinking	Personal attributes and skills	Mean interview score
Mean (SD)	3.16 (.63)	2.55 (.83)	2.65 (.89)	2.66 (.89)	2.67 (.92)	2.92 (.88)	2.77 (.70)
Correlations with SJT score	.24*	.31**	.21*	.21*	.21*	.21*	.29**

Table 3. Correlations Between Interview Scores and SJT Total Score

	Interview domains						
	Ability to communicate	Pedagogical & subject knowledge	Reflections on experience	Understanding of education	Quality of thinking	Personal attributes and skills	Mean interview score
Mean (SD)	3.16 (.63)	2.55 (.83)	2.65 (.89)	2.66 (.89)	2.67 (.92)	2.92 (.88)	2.77 (.70)
Correlations with SJT score	.24*	.31**	.21*	.21*	.21*	.21*	.29**

Note. *N* = 108. **p* < .05. ***p* < .01.

3. DISCUSSION

Developing evidence-supported ITE selection practices is one approach to improving system-wide educational outcomes. In this proof-of-concept study, we presented the development and initial validation of a test for selection into primary ITE programs. The novel contribution of this article is that we show, as far as we know, the development of the first SJT-based selection test for primary teacher education programs, and although the results are encouraging, they represent the first step of many in a move to develop an operational selection tool. The results from the study suggest that the SJT methodology shows potential for selection purposes, with evidence of reliability, validity, and a positive response (e.g., perceived fairness) from ITE candidates.

We examined four research questions in this study. In response to the first research question (*Can a robust set of target attributes be established?*), three target attribute clusters were developed from a systematic inductive approach and endorsed by a diverse group of teachers and teacher educators. The three domains derived from the inductive development process used in our research have corollaries in other conceptual models of teacher effectiveness and teacher-student interactions. Pianta and Hamre's CLASS framework (2009) proposes three domains—emotional supports, classroom organization, and instructional supports—that can be mapped on to at least two of the inductively-derived domains in our model. Our domain of Empathy and Communication shares common ground with Pianta and Hamre's *emotional supports*, especially with the dimensions of *teacher sensitivity* and *regard for student perspectives*. Our domain of Organisation and Planning shares commonalities with *classroom organization,* with its dimensions of *behavior management* and *instructional learning formats*. Models of teacher effectiveness developed by other researchers, e.g., the *self-regulation skills* and *motivational characteristics* from the work of Kunter, Kleickmann, Klusmann, and Richter (2013) also share aspects of the domains developed in our model.

The inductive approach that we used, involving practicing teachers and teacher educators, was rigorous, and the target attributes were shown to be robust. However, further work is needed to expand the target attributes to include theory-derived (deductive) attributes that have been associated with teaching effectiveness, such as personality (Rockoff, Jacob, & Kane, 2011) and self-efficacy (Klassen & Durksen, 2014).

Our second and third research questions pertained to the psychometric properties of the proof-of-concept SJT. The psychometric results were acceptable, with a high level of reliability, a near-normal distribution, and significant empirical associations with interview criteria. Internal consistency reliability coefficients for SJTs are often low, partly because contextualised items (scenarios) tend to be complex and measure multiple constructs, even when they are designed to assess a particular attribute (Patterson et al., 2015).

The concurrent validity coefficient of $r = .29$ with overall interview score is encouraging for a proof-of-concept study and it is in line with fully developed SJTs (Christian et al., 2010). Further research will be needed to establish incremental validity of the SJT (i.e., what the SJT adds to selection decisions over-and-above other selection measures) and further work is needed to explore the predictive validity of the test using reliable and valid measures of teaching effectiveness (e.g., Pianta & Hamre, 2009).

Our fourth research question (*Do candidates view the SJT as fair and as a feasible selection method?*) was answered by candidates' generally positive responses to completing the SJT during selection. Candidates' perceptions of selection practices influence acceptance decisions, likelihood of litigation based on perceived unfairness of acceptance policies, and the academic reputation of the selecting institution. Previous research has shown that contextualised selection methods (e.g., SJTs) are perceived as being fairer than non-contextualised methods (e.g., personality tests; Bauer & Truxillo, 2006). Further steps to increase transparency might include providing candidates with information about how the test was developed and validated, and how SJT scores would be integrated into the selection process (e.g., the amount of weight an SJT score would carry in the overall selection process).

3.1. How an SJT might be used for selection into ITE programs

For live selection, admissions committees could use the SJT test in two ways. First, the test could be used for initial screening of non-academic attributes before candidates are invited to an expensive and time-consuming assessment centre or face-to-face interview day. The scoring of the SJT provides an overall score that can be weighted along with other assessment criteria, such as academic records, letters of reference, and interview scores, to produce a screening cut-off score. Most ITE programs already screen for academic attributes (e.g., review of academic transcripts) before inviting applicants to interviews; the SJT could be offered on site or at invigilated test centres for screening of non-academic attributes. SJTs could also be used in place of interviews, providing an efficient, economical, and arguably more valid assessment of non-academic attributes. Finally, SJTs could be used in addition to (or in combination with) currently used measures of non-academic attributes (e.g., letters of reference, interviews) as an additional source of data for decision making that might provide improvement in predicting who would most likely be most effective teachers.

Next Steps. The results from the proof-of-concept SJT for selection are encouraging, but more psychometric and conceptual work is needed before such a test could be used for 'live' selection. Further work includes the generation of more SJT items to populate an item bank. Item development is an expensive and time-consuming process that requires item-writers to interview experienced teachers (who have worked with novice teachers) about critical incidents in a teaching context. Nevertheless, it is important to create a larger pool of validated items to populate alternate test forms in order to combat coaching effects (Whetzel & McDaniel, 2009).

The current study showed evidence of concurrent validity, but predictive validity evidence is needed to provide additional information about the usefulness of the SJT for ITE selection. While there is a lack of predictive validity research for any teacher selection process (Goldhaber, Grout, & Huntington-Klein, 2014), most SJT research explores concurrent, not predictive validity (Campion et al., 2014). A next step in developing a wider evidence base will be to study the relationships between pre-service teacher's SJT scores at entry and at the end of the ITE program. Further research will examine the longer-term predictive validity of SJTs using measures of teacher effectiveness in professional practice. Such tools may include the CLASS observation system (Pianta & Hamre, 2009), which involves observations of teachers' classroom behaviors, and the Tripod Survey, which involves anonymous student ratings of teacher-student interaction quality and classroom climate, which was used in the *Measures of Effective Teaching* project (Kane & Staiger, 2012). CLASS and Tripod measures are well-researched teacher effectiveness tools that have been rigorously validated over the last decade.

A further step will be to examine the relative effectiveness of competing constructs and selection measures. Lievens & Patterson (2011) used structural equation modelling to estimate the relative influence of SJTs alongside two other variables in predicting supervisor ratings of medical trainees' performance. Results showed that all three variables were valid predictors of job performance, with SJTs showing incremental validity over the academic measures. Final validation of an SJT designed for ITE selection would test incremental validity over the academic and non-academic measures currently used for selection.

We used a bottom-up inductive approach by way of a critical incidents technique to develop the target attributes to base our test content on. Another approach used in SJT research is a theory-based or deductive approach (Campion et al., 2014), in which target attributes are based on existing theoretical models such as personality and motivation. Our research team is currently developing theory-based SJTs to assess motivation (e.g., self-efficacy) and personality as target attributes.

3.2. International research

Interest in developing evidence-led ITE selection methods is not unique to the UK, and research on identifying key factors related to success in ITE programs is being carried out in a range of international settings. One key question in our international projects on ITE selection is the extent to which teaching attributes identified in one context are endorsed in another national context. A key principle in developing selection methods internationally is to recognize that although some attributes of effective teachers may be universal, other attributes measured need to reflect local contexts (Lievens et al., 2015).

3.3. Limitations

The sample of participants in Step 9 (pilot study) was smaller than anticipated and less ethnically diverse than the overall population of teachers in the UK (97.5% White British in our sample versus 93% nationally). However, the gender balance of participants in our study was the same (80%) as the gender balance reported for teachers nationally (Department for Education, 2016). One stated advantage of using SJTs for selection—that they are less prone to inter-group differences than other selection methods such as cognitive tests (Whetzel & McDaniel, 2009)—was not tested in this study, and more diverse samples will be needed to establish inter-group profiles to further investigate the fairness of SJTs.

4. CONCLUSIONS

This study is the first to report the development of a proof-of-concept SJT to select candidates into ITE programs. The results should be interpreted cautiously, with a restricted sample involving concurrent validity data. A selection system needs to be robust, transparent, and perceived as fair by applicants, and built on evidence collected from multiple methods. In many contexts, cost-effectiveness is also an important factor in choosing selection tools: an SJT can be used as a screening tool to evaluate non-academic attributes alongside evaluation of academic attributes, thus reducing the time and cost involved in the selection process. In settings where large numbers of candidates apply for limited spaces, SJTs could be used in conjunction with other data (such as academic records) to select a reduced number of candidates for more intensive selection procedures such as face-to-face interviews. The intention of this proof-of-concept study was to show the feasibility of developing an SJT for selection into teacher education programs, but exactly how, when, and the extent to which this method might be used would be determined by local contexts and needs.

Funding

This work was supported by research funding from the European Research Council.

5. REFERENCES

Anderson, L. & Wilson, S. (1997) Critical incident technique. In D.L. Whetzel and G.R. Wheaton (eds.), *Applied Measurement Methods in Industrial Psychology*. Palo Alto, CA: Davies-Black.

Atteberry, A., Loeb, S., & Wyckoff, J. (2015). Do first impressions matter? Improvement in early career teacher effectiveness. *AERA Open*, 1(4), 1-23.

Barber, M. & Mourshed, M. (2007). *How the world's best performing school systems come out on top*. London: McKinsey & Company.

Bauer, T. N., & Truxillo, D. M. (2006). Applicant reactions to situational judgment tests: Research and related practical issues. In J. A. Weekley & R. E. Ployhart (Eds.) *Situational judgment tests: Theory, measurement, and application* (pp. 233-249). Mahwah, NJ: Lawrence Erlbaum Associates.

Bergman, M. E., Drasgow, F., Donovan, M. a., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, 14, 223-235.

Campion, M. C., Ployhart, R. E., & MacKenzie, W. I. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. *Human Performance*, 27, 283–310.

Casey, C. E., & Childs, R. A. (2007). Teacher education program admission criteria and what beginning teachers need to know to be successful teachers. *Canadian Journal of Educational Administration and Policy*, 67, 1-24.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, 63, 83-117.

Department for Education (2016). Statistics at DfE. Retrieved from: https://www.gov.uk/government/organisations/department-for-education/about/statistics

Elliott, J. G., Stemler, S. E., Sternberg, R. J., Grigorenko, E. L., & Hoffman, N. (2011). The socially skilled teacher and the development of tacit knowledge. *British Educational Research Journal*, 37, 88-103.

Goldhaber, D., Grout, C., & Huntington-Klein, N. (2014). *Screen twice, cut once: Assessing the predictive validity of teacher selection tools*. CEDR Working Paper No. 2014-9: University of Washington, Seattle, WA.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition attitudes, self-esteem, and stereotypes. *Psychological Review*, 102, 4-27.

Hanushek, E. A., & Rivkin, S. G. (2011). The Distribution of Teacher Quality and Implications for Policy. *Annual Review of Economics*, 4, 131-157.

Hobson, A. J., Ashby, P., McIntyre, J., & Malderez, A. (2010). *International approaches to teacher selection and recruitment*. OECD Education Working Paper No.47: Organisation for Economic Co-operation and Development.

Hooper, A. C., Jackson, H. L., & Motowidlo, S. J. (2004). *Situational judgment measures of* *personality and work-relevant performance*. Paper presented at the 112th annual meeting of the American Psychological Association, Honolulu, HI.

Johnson, R. E., & Saboe, K. N. (2011). Measuring implicit traits in organizational research: Development of an indirect measure of employee implicit self-concept. *Organizational Research Methods*, 14, 530-547.

Kane, T. J., & Staiger, D. O. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains*. Research Paper. MET Project. Bill & Melinda Gates Foundation.

Klassen, R. M., & Dolan, R. (2015, September). *Selection for teacher education in the UK and the Republic of Ireland: A proposal for innovation*. Presented at the meeting of the European Conference on Educational Research, Budapest, Hungary.

Klassen, R. M., & Durksen, T. L. (2014). Weekly self-efficacy and work stress during the final teaching practicum: A mixed methods study. *Learning and Instruction*, 33, 158-169.

Klassen, R.M., Durksen, T.L., Rowett, E., & Patterson, F. (2014). Applicant reactions to a situational judgment test used for selection into initial teacher training. *International Journal of Educational Psychology*, 3, 104-125.

Knight, F. B. (1922). Qualities related to success in elementary school teaching. *The Journal of Educational Research*, 5, 207-216.

Kunter, M., Kleickmann, T., Klusmann, U., & Richter, D. (2013). The development of teachers' professional competence. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers* (pp. 63-77). New York, NY: Springer.

Lievens, F., Corstjens, J., Sorrel, M. A., Abad, F. J., Olea, J., & Ponsoda, V. (2015). The cross-cultural transportability of situational judgment tests: How does a US-based integrity situational judgment test fare in Spain. *International Journal of Selection and Assessment*, 23, 361-372.

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. *Journal of Applied Psychology*, 96, 927-940.

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

Lievens, F., & Sackett, P. R. (2012). The validity of interpersonal skills assessment via situational judgment tests for predicting academic succss and job performance. *Journal of Applied Psychology*, 97, 460-468.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel psychology*, 60, 63-91.

McDaniel, M. A., Whetzel, D., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, 79, 599-616.

Mikitovics, A., & Crehan, K. D. (2002). Pre-professional skills test scores as college of education admissions criteria. *The Journal of Educational Research*, 95, 215-223.

Motowidlo, S. J., & Beier, M. E. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. *Journal of Applied Psychology*, 95, 321-333.

Patterson, F., Ashworth, V., Mehra, S., & Falcon, H. (2012). Could situational judgement tests be used for selection into dental foundation training? *British Dental Journal*, 213, 23–26.

Patterson, F., Zibarras, L., & Ashworth, V. (2016). Situational judgement tests in medical education and training: Research, theory and practice. AMEE Guide No. 100. *Medical Teacher*, 38, 3-17.

Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., & Cleland, J. (2016). How effective are selection methods in medical education and training? Evidence from a systematic review. Medical Education, 50, 36–60.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, 38, 109-119.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education*, 6, 43-74.

Sahlberg, P. (2014). *Finnish lessons 2.0: What can the world learn from educational change in Finland?* New York, NY: Teachers College Press.

Schmitt, N., & Chan, D. (2006). Situational judgment tests: Method or construct? In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 135-155). Mahwah, NJ: Erlbaum.

Sclafani, S. K. (2015). Singapore chooses teachers carefully. Phi Delta Kappan, 97(3), 8-13.

Shultz, M. M., & Zedeck, S. (2012). Admission to Law School: New Measures. *Educational Psychologist*, 47, 51-65.

Staiger, D. O., & Kane, T. J. (2015). Making Decisions with Imprecise Performance Measures. In T. Kane, K. Kerr, R. Pianta, & Bill and Melinda Gates Foundation (Eds.), *Designing Teacher Evaluation Systems* (pp. 144-169). San Francisco, CA: Jossey-Bass.

Staiger, D. O., & Rockoff, J. E. (2010). Searching for effective teachers with imperfect information. *Journal of Economic Perspectives*, 24(3), 97-118.

Thomson, D., Cummings, E., Ferguson, A. K., Moizumi, E. M., Sher, Y., Wang, X., Broad, K., & Childs, R. A. (2011). A role for research in initial teacher education admissions: A case study from one Canadian university. *Canadian Journal of Educational Administration and Policy*, 67, 1–24

Walker, H. J., Bauer, T. N., Cole, M. S., Bernerth, J. B., Feild, H. S., & Short, J. C. (2013). Is this how I will be treated? Reducing uncertainty through recruitment interactions. *Academy of Management Journal*, 56, 1325-1347.

Watt, H. M. G., Richardson, P. W., & Wilkins, K. (2014). Profiles of professional engagement and career development aspirations among USA preservice teachers. *International Journal of Educational Research*, 65, 23-40.

Whetzel, D., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review*, 19, 188-202.