Teaching and Teacher Education 87 (2020) 102949

Contents lists available at ScienceDirect

Teaching and Teacher Education

journal homepage: www.elsevier.com/locate/tate

Can we improve how we screen applicants for initial teacher education?

Robert M. Klassen*, Lisa E. Kim, Jade V. Rushby, Lisa Bardach

University of York, UK

HIGHLIGHTS

- A test for screening applicants to teacher education was developed.
- Internal consistency of the test was acceptable.
- The test was significantly correlated with interview performance.
- The test was more predictive of interview performance than current screening methods.
- High scorers on the test performed better at interview than low scorers.

ARTICLE INFO

Article history:
Received 14 December 2018
Received in revised form
18 May 2019
Accepted 30 September 2019
Available online xxx

ABSTRACT

Identifying the best possible candidates for initial teacher education (ITE) programs is one of the first steps in building a strong teacher workforce. We report three phases of development and testing of a contextualized teaching-focused situational judgment test (SJT) designed to screen applicants at a large and competitive ITE program in the U.K. Results showed that the SJT was a reliable and predictive tool that enhanced existing screening methods. We suggest that using state-of-the art methods to help make admissions decisions could improve the reliability, validity, and fairness of selection into ITE.

1. Introduction

One of the first steps in the development of an effective teacher workforce is to identify applicants who first, are likely to succeed in an initial teacher education (ITE) program, and second, are likely to experience success as practicing teachers. Evidence for individual differences in the developmental trajectory of teachers is persuasive (Atteberry, Loeb, & Wyckoff, 2015; Chetty, Friedman, & Rockoff, 2014; Hanushek & Rivkin, 2012; Xu, Özek, & Hansen, 2015), with both cognitive attributes (e.g., academic ability, subject knowledge, pedagogical knowledge) and non-cognitive attributes (e.g., interpersonal skills, personality, and motivation) hypothesized to contribute to these differences (Klassen & Tze, 2014; Rockoff, Jacob, Kane, & Staiger, 2011). Collecting robust data on applicants' *cognitive* attributes at the point of selection into ITE is comparatively straightforward, with academic records from university and secondary school widely available to selectors, and tests of academic ability and subject knowledge available from a wide range of sources (e.g., ETS Praxis, n.d.).

Assessing applicants' *non-cognitive* attributes in a way that is reliable, predictive, fair (more objective and less prone to interviewer bias) and efficient (in terms of time and cost) is more difficult. The importance of teachers' non-cognitive attributes can be traced to the very beginning stages of training and professional practice (Bastian, McCord, Marks, & Carpenter, 2017; Watt, Richardson, & Wilkins, 2014), but identifying and assessing these attributes at the point of selection has proven to be methodologically challenging and time-consuming, with weak relations between selection methods and subsequent teacher effectiveness (Klassen & Kim, 2019; Bieri & Schuler, 2011; Rimm-Kaufman & Hamre, 2010). The assessment of non-cognitive attributes for selection into ITE is not often critically examined, but when it is, results show low predictive validity (e.g., Casey & Childs, 2011; Klassen & Kim, 2019). The novel contribution of this article is that we describe (to our knowledge, for the first time) how a methodology used to assess non-cognitive attributes for selection in other professional fields (e.g., in business or medical education) can be

* Corresponding author. Psychology in Education Research Centre, University of York, York YO10 5DD, UK.
E-mail address: robert.klassen@york.ac.uk (R.M. Klassen).

https://doi.org/10.1016/j.tate.2019.102949

used as a reliable, valid, fairer, and efficient screening measure for applicants to teacher education, with the potential to improve the teacher workforce.

1.1. Selection methods for teacher education programs

Selection methods for ITE (also known as teacher preparation programs or preservice teacher education) are designed to evaluate the potential for program and professional suitability based on an assessment of personal characteristics (cognitive and non-cognitive attributes) and background factors, such as academic qualifications and relevant experiences. The urgency of the need to develop reliable and valid ITE selection processes varies across countries, ITE programs, and subject areas (Davies et al., 2016; Greenberg, Walsh, & McKee, 2015; Ingvarson & Rowley, 2017), with some programs facing a shortage of applicants and other programs needing to make difficult decisions about which applicants to select. However, even in settings where concerns about recruitment outweigh concerns about selection, using reliable, valid, fair, and efficient methods to assess applicants' cognitive and non-cognitive attributes can lead to a more robust understanding of the factors that influence teacher development at 'Year 0' of a professional career.

Personal characteristics: Cognitive and non-cognitive attributes. In Kunter et al.'s COACTIV (Cognitive Activation) model of professional competence (Kunter, Kleickmann, Klusmann, & Richter, 2013), teachers' personal characteristics play an important, but not exclusive, role in influencing the development of professional competence. Professional competence is also influenced by environmental factors, such as the quality of learning opportunities (e.g., subject and pedagogical support during teacher education), and other diverse factors (e.g., quality of in-school mentoring and support). However, at the point of selection into ITE programs, it is applicants' personal characteristics that are under closest scrutiny. Cognitive attributes are evaluated during selection by proxies such as academic transcripts and/or scores from tests of academic aptitude (e.g., Praxis). An assumption about the centrality of cognitive attributes for teacher selection is sometimes made whereby the best teachers are believed to be those who are "smart enough and thoughtful enough to figure out the nuances of teaching" (Kennedy, Ahn, & Choi, 2008, p. 1248). However, the evidence linking teachers' cognitive abilities and classroom effectiveness is equivocal (Harris & Rutledge, 2010), with review studies showing either small associations (e.g., Wayne & Youngs, 2003), or non-significant or even *negative* associations between cognitive attributes and effectiveness (e.g., Bardach & Klassen, 2019).

Assessment of non-cognitive attributes for ITE selection is also challenging, with little agreement about what to measure and which methods to use. The range of non-cognitive attributes targeted by ITE programs at selection is very wide, including confidence, integrity, resilience, motivation, the 'X factor' (Davies et al., 2016), and inter-personal skills (e.g., Donaldson, 2011). A recent cross-national study (Klassen et al., 2018) on the desired non-cognitive attributes of preservice and novice teachers found that some attributes (*empathy, organization,* and *adaptability*), were universally endorsed, whereas other attributes (*fostering of community, autonomy,* and *enthusiasm*), were associated with particular settings. The methods most frequently used in ITE to assess non-cognitive attributes (Klassen & Durksen, 2015)—letters of reference, personal statements, motivational essays, and traditional interviews—show modest evidence supporting their use, and furthermore, may be prone to selectors' conscious or unconscious biases (e.g., Mason & Schroeder, 2014; Patterson et al., 2016).

We agree with Kunter's COACTIV model that teacher effectiveness (i.e., competence) is malleable (Kunter et al., 2013), but we suggest that personal characteristics evaluated at the point of selection into ITE influence future effectiveness. ITE programs strive to choose applicants who are higher, rather than lower, in certain cognitive attributes (e.g., reasoning abilities, subject knowledge), and non-cognitive attributes (e.g., conscientiousness, self-regulation, adaptability, and empathy), while acknowledging that personal characteristics will interact with environmental factors and learning opportunities over the course of a career.

Current approaches to evaluate personal characteristics. Current methods used to evaluate prospective teachers' cognitive and non-cognitive attributes are not very convincing. A recent meta-analysis (Klassen & Kim, 2019) examined research that reported the predictive validity of selection methods for prospective teachers, both for selection into employment and into ITE programs, with outcomes defined as 'objective' measures of the quality of teaching (i.e., classroom observation and student achievement gains measures, but not self-report). The results showed an overall effect size of $r = 0.12$ ($df = 31$, $p < .001$) across 32 studies, with the effect size for non-cognitive predictors ($r = 0.10$) smaller than for cognitive predictors ($r = 0.13$). The effect size for selection into ITE programs was $r = 0.14$, with all effect sizes in the meta-analysis smaller than those typically found for selection in other fields (e.g., mean r of 0.26; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). There are at least three interpretations of the low effect sizes found in teacher selection research: teaching is uniquely complex (e.g., see Maulana, Helms-Lorenz, & van de Grift, 2015 for a study on the relative complexity of teaching domains), the methods used to measure teaching outcomes are poor, or the methods used for ITE selection are poor, and do not reflect the research and development that has characterized other fields.

Selection methods from other fields. Research from organizational psychology and health-related fields can provide new ideas for ITE selection. A recent review of selection methods for medical education (Patterson et al., 2016) found that for the measurement of cognitive attributes, academic records and performance on aptitude tests (e.g., medical entrance exams such as the American-based Medical College Admissions Test [MCAT] and UK-based University Clinical Aptitude Test [UCAT]) provide some predictive power for medical school performance and success in professional practice. For the assessment of non-cognitive attributes, structured interviews, assessment centers and situational judgment tests (SJTs) were judged to be more effective (higher predictive validity) and fairer (less prone to selectors' unconscious biases) than reference letters, personal statements, and unstructured interviews.

Although structured interviews may be associated with future performance, they can be time-consuming, and thus, expensive, to conduct. Metzger and Wu (2008) assessed the Teacher Perceiver Interview and its online format, TeacherInsight, which is a commercial interview method used by schools to select teachers in 15% of school districts in the US. Initial start-up costs, annual fees, and training were found to be expensive, and the interviews took between 40 min (TeacherInsight) to 2 h (Teacher Perceiver Interview) to conduct, and validity was found to be moderate (0.28) (Metzger & Wu, 2008). Furthermore, interviews have been found to be subject to interviewer bias (Davison & Burke, 2000) and may not accurately measure target attributes but other personal factors such as the likability of the applicant (Schumacher, Grigsby, & Vesey, 2013). Assessment centers (called 'interview days' in some ITE contexts), which use a modular approach to selection, incorporating a series of tests, exercises, and structured interviews to assess non-cognitive attributes, are frequently used in business settings, and show higher levels of predictive validity (Lievens & Sackett, 2017). However, assessment centers are expensive to conduct: for large-scale selection, a screening process may be

necessary to manage the numbers of applicants invited to interview days. Thus, new methods, such as situational judgment tests (SJTs), that are efficient, reliable, and valid may be useful in the selection process for teacher education.

1.2. Situational Judgment Tests (SJTs)

SJTs are a measurement method in which test-takers read a brief, context-rich, 'real-life' scenario, and then provide an assessment of the appropriateness of a range of responses. A typical SJT scenario presents a challenging social situation (for example, in a classroom), followed by the question *What should (or would) you do?* and a series of response options. The scenarios and responses are designed to assess an applicant's procedural knowledge and situational judgment. Test-takers with particular knowledge and experience in a setting, e.g., ITE applicants who have worked in schools, may draw on that experience when responding, but SJTs also assess a candidate's non-specific situational judgment, independent of experience in a specific professional context (Melchers & Kleinmann, 2016). Situational judgment refers to the ability to perceive and interpret an ambiguous social situation, with research showing that situational judgment is a valid predictor of a range of work performance outcomes (Rockstuhl, Ang, Ng, Lievens, & Van Dyne, 2015).

At a theoretical level, SJTs are hypothesized to assess a test-taker's *implicit trait policies*, or evaluation of the costs and benefits of particular courses of action (Motowidlo, Hooper, & Jackson, 2006). Implicit trait policies are grounded in a person's underlying traits and beliefs that have developed from fundamental socialization processes (i.e., from family, schooling, peers, etc.). For example, in an SJT scenario targeting the personality trait of agreeableness, people will rely on their underlying traits and beliefs to make decisions about the most appropriate level of cooperation or competitiveness required for effective action in a given situation (Lievens & Motowidlo, 2016). Implicit trait policies inform the internal decision-making about the best course of action in a challenging situation.

The popularity of SJTs among organizational psychologists and selection panels is based on their relatively high levels of predictive validity, their ease-of-use, cost effectiveness once developed, and generally favorable applicant reactions (e.g., Klassen et al., 2017; Lievens & Motowidlo, 2016). In some ways, SJTs are similar to structured interview questions used in many ITE selection settings (*Tell us what you would do if …*), but with the added advantages of gathering multiple samples of applicant judgment and a more systematic and objective scoring system. In a face-to-face interview, applicants might be asked to respond to one or two contextualized scenarios, whereas in an SJT, responses to a high number of scenarios (25 scenarios in Phase 2 and 3 of this study) can be collected in a relatively brief period of time. In addition, SJTs use a structured scoring system that is applied in the same way across applicants, with less possibility that the unconscious biases which influence face-to-face interview scoring (based on age, race, sexuality, class, gender, social background, and physical attractiveness), will affect the scoring (Cook, 2009).

In order to develop realistic, contextualized SJT scenarios, test developers must enlist the help of subject matter experts (in this case, expert teachers) who are familiar with the professional challenges in the field of interest (Patterson et al., 2016). Because SJTs are constructed from complex, multi-faceted, real-life scenarios, they are often heterogeneous at the item level, mapping on to multiple constructs (McDaniel, List, & Kepes, 2016). The factor structure of SJTs can be ambiguous, with exploratory factor analyses typically revealing multiple uninterpretable factors (Lievens, Peeters, & Schollaert, 2008), or a unidimensional structure rather than the hypothesized multidimensional structure (e.g, Fröhlich, Kahmann, & Kadmon, 2017). In spite of the psychometric challenges and expense of test development, SJTs are increasingly used to assess non-cognitive attributes for selecting candidates for training in a wide range of professional fields because they are good predictors of work-related outcomes (Buyse & Lievens, 2011; Taylor, Mehra, Elley, Patterson, & Cousans, 2016). Longitudinal validity studies show that SJTs developed to test medical school applicants' non-cognitive attributes reliably predicted professional effectiveness several years after the selection process (Lievens & Sackett, 2012), and SJTs were rated as more effective for screening applicants for medical education than aptitude tests, personal statements, reference letters, and personality tests (Patterson et al., 2016). Research on SJTs for selection has been conducted in multiple professional fields, but use of the methodology for selection into ITE programs has received only modest attention (Klassen & Kim, 2017).

1.3. Current study

Developing SJTs for ITE selection requires multiple steps to ensure that the tests are reliable, valid, and accurately reflect the target educational context. In this article, we report the development and testing of an online SJT designed to screen applicants for invitation to an ITE interview day (or 'Assessment Center'). The ITE program in this study is large and competitive, based in London, and draws applicants from all over the UK. This ITE program uses a three-step selection process: (1) online eligibility checks (i.e., checking qualifications), followed by (2) online screening tests as a sift to assess suitability for the program, leading to (3) invitation to an on-site interview day involving multiple selection activities. The SJTs were developed for use as part of the online screening tests in step (b). All stages of the research (i.e., development and administration) were reviewed and approved by the first author's university ethics review board and by the selection and recruitment team at the ITE site (the authors of the current article are not formally affiliated with the ITE program in question, and were not involved in making selection decisions).

Three phases of development and testing were conducted as part of this study. Phase 1 involved the development of the SJT content, Phase 2 included the administration of the initial SJT prototype, and in Phase 3, the revised SJT was administered to ITE program applicants alongside other online screening tests. The primary research questions are:

1. What are the psychometric properties of a teacher selection SJT (reliability, concurrent and predictive validity, factor structure, and statistical relationships with other screening tools and interview day activities)?
2. Does the SJT provide incremental predictive validity beyond screening methods currently used?
3. Do high scorers on the SJT fare better on interview day activities than low scorers on the SJT?

2. Phase 1: development of a construct-informed SJT for screening ITE applicants

Participants in the development phase of the SJT were 19 expert teachers (13 females, 6 males) who were involved in administering ITE selection activities. We defined 'expert' as (a) > 5 years' experience as a teacher, (b) recent experience as an interviewer on the selection process, or (c) recent experience with systematic observation of novice teachers. The expert teachers worked with the research team to identify key attributes, develop the test specification, develop and review test items, and set the scoring key for

the SJT. The development activities were conducted between 2015 and 2017.

Identifying foundation attributes. The process for identifying the foundation attributes on which the SJT was built followed an integrated inductive and deductive approach (e.g., Guenole, Chernyshenko, & Weekly, 2017; Schubert et al., 2008; Weekley, Ployhart, & Holtz, 2006). The majority of SJTs are developed using an inductive approach where the key attributes are identified during the content development process (Campion, Ployhart, & MacKenzie, 2014). In this approach, researchers work with experts to identify critical incidents related to the field of interest, and subsequently assign inductive categories to the content. In contrast, SJTs developed using a deductive approach identify target attributes before the content development process and develop content that represents the targeted attributes (e.g., Guenole et al., 2017). We used an integrated 'construct-informed' (or construct-driven; Lievens, 2017) inductive and deductive approach, in which three non-cognitive attributes emerged from a series of interviews with experts (i.e., 'bottom-up'), and three non-cognitive attributes were targeted *a priori* based on existing theories ('top-down'). The three inductive attributes—adaptability, organization, and empathy—were previously developed through a multi-step inductive process reported in Klassen & Tze, 2014, Klassen et al., 2017. The three deductive attributes—conscientiousness, growth mindset, and emotion regulation—were chosen through a review of relevant literature and through a series of discussions with ITE program staff. Conscientiousness was chosen as a target attribute because it has been shown to be one of the Big Five personality domains most related to teacher effectiveness (e.g., Kim, Dar-Nimrod, & MacCann, 2018; Kim & MacCann, 2018); growth mindset was chosen because of the increasing recognition that teachers' beliefs influence how students perceive their learning (e.g., Seaton, 2017), and emotion regulation was chosen because teacher emotions and emotion regulation are related to a range of important teaching-related outcomes (e.g., Chang, 2013; Sutton, 2004; Taxer & Gross, 2018). The six non-cognitive attributes were used as a guide in the creation of scenarios of the SJT.

Test specification. The test specification—(a) purpose of the test, (b) test content, (c) item types, (d) response formats used, and (e) desired length of the test—was developed by the research team and key members of the ITE program. The (a) purpose of the test was to provide an initial online screening of applicants to an ITE program. The (b) test content was developed to evaluate applicant judgment related to the targeted non-cognitive attributes (i.e., conscientiousness, growth mindset, emotion regulation, adaptability, organization, and empathy). The (c) item type was determined to be scenarios of challenging classroom situations followed by response options. The (d) response format was a 4-point Likert rating scale indicating degree of appropriateness ('inappropriate' to 'appropriate'). As recommended by Whetzel and McDaniel (2016), we used 'should' instructions rather than 'would' instructions (i.e., *What should you do in this situation?*) in order to reduce candidate faking, since with this format all respondents have the same goal: to identify the best course of action in a particular context. The (e) desired length of the test was determined to be 30 min or less.

Item development. The 19 expert teachers were interviewed by three members of the research team using a critical incident approach (e.g., Buyse & Lievens, 2011). Participants in this phase were given the following written instructions (abridged) one week before individual meetings:

We are developing a tool that focuses on evidence-based attributes shown to be associated with successful teaching (definitions were provided for conscientiousness, emotion regulation, and mindset). During our face-to-face conversation, we will ask you to share two scenarios or incidents that are related to these attributes. These scenarios should relate to situations novice teachers might be expected to deal with, and the incident must relate to one (or more) of the six target attributes. We will also ask you to provide potential responses to the scenario and to rate the appropriateness of each response.

The interviews were scheduled to last 45 min, and the researchers recorded scenarios and responses on an item development template. A total of 48 items was generated over two days of interviewing, with most items accompanied by five response options (range: 4–8 response options, with the goal of identifying four 'good' options at the review panel).

Review panel. The 48 items created in the item development phase were initially reviewed by the research team, who edited the scenarios to eliminate errors, inappropriate and redundant items, and items that did not clearly map onto the target attributes. The SJT items were administered to a review panel consisting of teachers from the ITE program, who completed the test (and provided additional comments). The initial scoring key was developed using a mode consensus approach (De Leng et al., 2017; Weng, Yang, Lievens, & McDaniel, 2018), with item response options reduced to four. Items showing a high degree of consensus were retained, whereas items with a low degree of consensus were set aside for further development. The final 25-item SJT included items that reflected the six target attributes, but the distribution was not equally divided among the attributes: emotion regulation 10 items, conscientiousness 6 items, growth mindset 3 items, empathy 3 items, adaptability 2 items, and organization 1 item.

Format and scoring. The SJTs were designed to be delivered online as part of the initial screening tests completed by all applicants to the ITE program. The scoring key for the SJT was developed using a hybrid scoring approach where expert-based scoring was used to set the initial key, but scoring key adjustments were made empirically (Bergman, Donovan, Drasgow, Henning, & Juraska, 2006), based on the review panel expert ratings. Item scores were calculated using a distance-measure approach, where a score is calculated based on distance from teacher-determined correct score (with three points for a correct response, two points for a response one position away from correct, one point for a score two positions away from correct, and zero for a response three positions away from correct). Thus, the maximum total score was 300 (25 scenarios, 4 response options x 3 maximum points for each response).

3. Phase 2: administration of a prototype SJT to ITE program applicants

The online screening process for this program runs on a near-continuous basis, with on-site interview days scheduled throughout the year. In this program, applicants for primary and secondary teacher training complete the same application process, and there is no differentiation in the process or activities used for selection. Data from the prototype SJT reported in Phase 2 were collected in 2017 and early 2018. The SJT was administered alongside the established ITE-developed screening process, but was not used for selection decisions.

Participants. The 3341 online applicants were 64.1% female, with a mean age of 26.49 years ($SD = 13.10$). Eligibility for free school meals (FSM) during their school years was used as a proxy for applicants' socio-economic background (e.g., Ilie, Sutherland, & Vignoles, 2017): 23.9% of applicants reported eligibility for FSM during their school years (in comparison, approximately 18% of UK primary school students were eligible for FSM in 2013; 13.7% in 2018; Department for Education, 2018).

Procedure. The online screening phase is designed to select candidates for the on-site interview day.

Online screening process. Applicants completed three tasks for the online application: after successful completion of an eligibility check, applicants were asked to respond to three competency-based written questions, and to complete the SJT. The eligibility check ensured that applicants had a relevant previous degree and were eligible for teacher training in the UK: acceptable A-level exam results in relevant subjects (usually taken at age 18), a grade 'C' or equivalent in General Certificate of Secondary Education (GCSE) exams in mathematics and English (usually taken at age 16), and an undergraduate degree (at level 2:1 or better) in a relevant teaching subject. Applicants completed the written tasks and SJT at their convenience on the device of their choice (computer 87%, tablet 1%, and phone 8%; with 4% not reported).

The three competency-based questions consisted of 250-word free responses (i.e., short essays) to prompts on (a) understanding of education and motivation for teaching, (b) leadership potential, and (c) a third competency chosen by the candidate (not included in these analyses due to the varied nature of the topics chosen, e.g., problem-solving, interaction, resilience). The score for each question was calculated as the mean score from two raters using an 8-point scale for each question.

Decisions for invitation to the interview day were based on the scores from the written questions alongside a review of academic qualifications and other contextual recruitment information.

Prototype situational judgment test. The prototype SJT consisted of 25 items delivered after completion of the written questions during the online screening session. The test was prefaced with the following instructions: *In this test, you are presented with scenarios that teachers encounter. Rate the appropriateness of each of the options in terms of what a (beginning teacher) should do (Inappropriate, Somewhat inappropriate, Somewhat appropriate, Appropriate), given the circumstances described in the scenario. There can be tied rankings, i.e., you can give multiple responses the same rating. Your rating on one option is independent from your ratings on the other options. For the test there are 25 questions, which should take 30 min to complete. Good luck!* Applicants were informed that the SJT data would be used strictly for research purposes.

Applicants were not given a time limit for SJT completion, and applicants were not proctored during the test, i.e., they completed the test at the place and time, and on the device, of their choosing. Fig. 1 provides a sample SJT item (similar, but not identical to items administered as part of the screening process).

Interview day. Applicants who scored above a certain threshold and who met other selection criteria (e.g., positive evaluation of related experiences, relevance of teaching subject) were invited to attend a day-long interview day, held on a rolling basis throughout the year, but typically applicants attended interview days about one month after the screening tasks were completed. In total, 831 out of 3341 applicants (24.9%) were invited to attend the Phase 2 interview day. Activities at the interview day included a competency-based 1-1 interview, a group activity centered around a case study, and a 7-min sample teaching demonstration. Each of the three activities was scored out of 40, with five competencies (*understanding and motivation, leadership, planning and organization, problem solving,* and *resilience*), each scored out of 8. Final decisions about acceptance to the program were based on interview day scores plus a consideration of other relevant factors (e.g., teaching subject area).

Analysis. The analysis of the prototype SJT included reliability analysis, correlation coefficients with screening and interview day tasks, and analysis of individual items of the SJT.

Phase 2 results. Brief summary results from analysis of the prototype SJT are presented in this section, with a more detailed analysis of the revised SJT presented in phase 3 results. The internal consistency (coefficient alpha) of the prototype SJT was 0.69, and bivariate correlations with the screening tests were $r = 0.28$, $p < .001$ with understanding and motivation score, and $r = 0.25$, $p < .001$ with the leadership score. Correlation of the prototype SJT with interview day scores were $r = .07$, $p = .01$ with interview, $r = 0.10$, $p = .001$ with group case study, $r = 0.10$, $p = .001$ with the sample teaching scores, and $r = 0.13$, $p = .001$ with total interview score.

4. Phase 3: administration of revised version of SJT to ITE applicants

In Phase 3, we first reviewed and refined the SJT content based on item analysis, and then administered the revised version of the SJT to applicants who applied to the ITE program during a 3-month period in mid-2018.

SJT revision. Scenarios and response options were reviewed and refined by the research team based on (a) item difficulty (i.e., proportion of correct response at the item level), (b) item discrimination (i.e., item-total correlations), (c) item correlations with Phase 2 interview day activities, and (d) response scoring patterns. The review process identified five items that were deemed to benefit from revision. For example, items with too-high or too-low item difficulty (i.e., did not discriminate among test-takers) or ambiguous response scoring patterns were improved by revising response options to increase clarity. Three members of the research team (two of whom had teaching experience) worked together to identify problematic items and to revise content, with a consensus-building approach to resolve differences. At the end of the revision process, the revised SJT was uploaded to the ITE application website and released for completion by new applicants.

Participants. Participants in Phase 3 were 587 applicants (61.7% female; 23.0% of whom self-reported as eligible for FSM as primary/secondary students), who completed the revised SJT as part of the screening process. Of the 587 applicants, 97 (16.5%) were invited to attend the interview, based on the screening criteria.

Procedure. The procedure for Phase 3 was identical to the procedure described in Phase 2.

Analysis. Analysis of the data comprised a descriptive analysis of means, range, and standard deviations for the key variables; assessment of online screening tasks for mean differences by gender and SES (defined as eligibility for FSM); ANCOVA to examine differences on interview performance by SJT scoring group; analysis of bivariate associations between SJT scores and key variables; and hierarchical multiple regression showing the contribution of screening scores to prediction of interview day activities.

Phase 3 results. The results from the revised SJT were assessed for reliability, group differences, factor structure, and associations with screening and interview day activities. An analysis of the five revised items showed improvements in item-total correlations and bivariate relationships with screening and interview day scores.

Internal consistency of the revised SJT was calculated using Cronbach's alpha, with a reliability coefficient of 0.78. In Table 1, we present the means, ranges, and standard deviations for the relationship between the SJT, screening scores (the two 'fixed' screening written questions, plus total screening score), and interview day scores (for individual interview, group case study, sample teaching demonstration, and interview day score). Fig. 2 shows the distribution of SJT scores, showing a negative skewness and a leptokurtic pattern, with scores clustering near the mean.

In order to test the factor structure of the revised SJT, we conducted a confirmatory factor analysis (CFA) using the target attribute structure from the SJT development, and followed this with a minimum average partial (MAP) test which has been suggested for construct-heterogeneous tests such as SJTs (Fröhlich et al., 2017).

You are teaching a Year 9 science class and the students are listening as you explain something on the whiteboard. At one point, you forget what you want to say next. As you pause, a girl in the front row laughs and says, “You’re useless!” but only loud enough so you and maybe some pupils next to her can hear.

Rate the appropriateness of each of the options in terms of what you should do as a first-year teacher:

Response options:

Quietly and firmly ask the student to leave the class in order to establish your authority in front of the other students *(Inappropriate)*

Quietly tell her that the comment was not acceptable and explain what consequence she will face *(somewhat appropriate)*

Ignore her, gather your thoughts and carry on (*somewhat appropriate)*

Turn and explain to the class what just happened, outline why it was inappropriate and what consequence she will face *(somewhat inappropriate)*

Fig. 1. Example item from situational judgment test.

Table 1
Descriptive statistics for SJT, screening test scores, and interview day scores.

	N	M (Range)	SD
SJT	587	240.75 (117.0–264.0)	13.69
Understanding and motivation (screening)	449	4.66 (1–8)	1.26
Leadership (screening)	449	4.76 (1–8)	1.32
Total screening	449	14.04 (3–22)	3.40
ID interview	97	28.07 (13–40)	5.33
ID case study	97	25.25 (1–39)	6.36
ID sample teaching	97	26.15 (0–41)	7.48
Total ID	97	79.47 (14–104)	15.54

Note. SJT = Situational Judgment Test. ID = Interview Day.

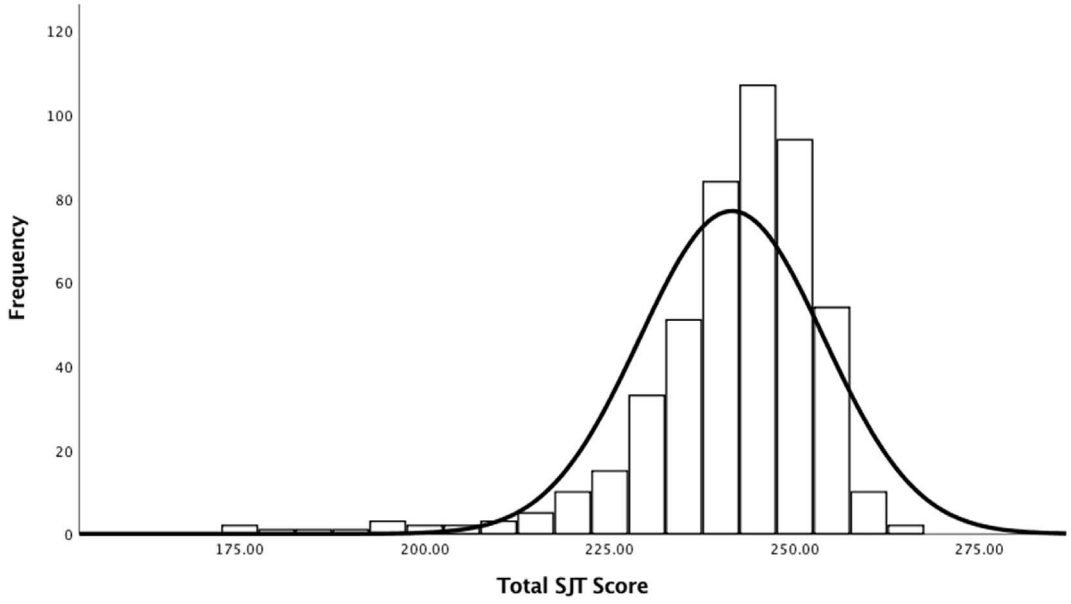

Fig. 2. Distribution of SJT scores.

Factor structure. The CFA was estimated using a Bayes estimator due to the advantages of Bayesian approaches, for example, the elimination of inadmissible parameters occurring when a maximum likelihood estimator is used (e.g., van de Schoot et al., 2014; also see Muthén & Asparouhov, 2012). Three CFA models were set up and compared. The Bayesian Markov chain Monte Carlo (MCMC) method was used to estimate our models. Eight chains were requested and a minimum number of 10,000 iterations were specified. Starting values were based on the maximum likelihood estimates of the model parameters, and Gelman-Rubin convergence statistics were used to check for convergence (Gelman & Rubin, 1992). For model comparisons, we relied on the Bayes Information Criteria (BIC) as a measure of the trade-off between model fit and complexity of the model, with lower BIC values indicating a better trade-off (e.g., van de Schoot et al., 2014). First, we conducted a one-factor CFA model, in which all SJTs loaded on a single first-order overall factor (Model A, BIC = 87018.43). Second, we modeled a five-factor model with five first-order factors for the five dimensions (Model B, BIC = 87449.41). As there was only one SJT item assessing "planning and organization", this item was added to the "empathy and communication" dimension (the scenario included elements of empathy and communication). Third, we conducted a model with five first-order factors that loaded on 1 s-order factor (Model C, BIC = 87538.12). A comparison of BICs showed an advantage of Model A (single, first-order factor) over the two alternative models.

The minimum average partial (MAP) test, recommended for use with construct-heterogeneous tests, is conducted by partialling each factor out of the correlation matrix and calculating a partial correlation matrix. The number of factors to be retained is indicated when the average squared partial correlation reaches a minimum level (Velicer, Eaton, & Fava, 2000). The results of the MAP test was consistent with the CFA and showed a single factor solution with an eigenvalue of 2.78, suggesting a uni-dimensional structure. Taken together, the two factor analyses showed that the SJT was best described through a single factor structure model.

Group differences. We compared the scores from screening for differences according to SES (measured as eligibility for free school meals [FSM]), gender, and type of device used to complete the screening tests. For the SJT, there were no significant differences for applicant SES measured by FSM, $F(1,390) = 0.15$, $p = .70$; $M_{HIGH} = 242.74$, $SD_{HIGH} = 14.26$; $M_{LOW} = 242.19$, $SD_{LOW} = 10.14$. For total screening score, there was a small but significant mean differences for SES as measured by applicant self-reported FSM, $F(1, 298) = 5.53$, $p = .02$, $\eta_p^2 = 0.02$; $M_{HIGH} = 14.57$, $SD_{HIGH} = 3.36$; $M_{LOW} = 13.60$, $SD_{LOW} = 3.35$.

Significant gender differences favoring females at screening were found for SJT scores $F(1,480) = 8.67$, $p = .003$, $\eta_p^2 = 0.02$ (considered 'small'; Cohen, 1988); $M_F = 242.6$, $SD_F = 10.01$; $M_M = 238.85$, $SD_M = 18.08$, but not for total screening score $F(1, 361) = 1.56$, $p = .21$, $M_F = 14.24$, $SD_F = 3.48$; $M_M = 13.77$, $SD_M = 3.58$.

Applicants who used a computer to complete the screening tests scored significantly higher, $F(3, 361) = 4.91$, $p = .002$ on the written responses to competency-based questions ($M = 14.34$, $SD = 3.45$) than applicants who used a phone ($M = 12.0$, $SD = 3.40$), but there were no significant differences on SJT scores by type of device used, $F(3, 480) = 1.64$, $p = .18$.

We divided participants into three equal groups (high, medium, and low scorers) according to their performance on the SJT. A one-way ANCOVA was conducted to determine the effect of SJT score on interview day scores ($M_{HIGH} = 87.82$, $M_{MED} = 75.47$, $M_{LOW} = 69.59$) controlling for gender and SES (using FSM as a proxy). There was a significant effect of SJT score on interview day score $F(2,62) = 12.37$, $p < .001$, $\eta_p^2 = 0.30$, considered a large effect size (Cohen, 1988).

Correlations and multiple regression. In Table 2, we report the bivariate correlations between SJT scores, screening scores, and interview day scores held about one month after screening tasks were completed (the time lag varied due to the rolling nature of both screening activities and interview days). The SJTs were significantly associated with screening tests (and total screening score), and with 4 out of 5 interview day scores, including total interview day score ($r = 0.46$, $p < .01$). SJTs were not significantly associated with the group case study activity ($r = 0.20$, $p =$ ns). Tables 3–5 report the association of SJT scores with individual competency scores within each of the interview day activities. In Table 3, SJT scores were significantly correlated with each of the five categories of the interview (*r*s ranging from 0.26 to 0.39), and more strongly correlated than the total screening score for each of the categories. In Table 4, we see that SJTs were only weakly associated with scores from the group case study activity (*r*s ranging from 0.06 to 0.25). In Table 5, we see that SJT scores were significantly associated with all categories of the sample teaching scores (*r*s ranging from 0.37 to 0.45), and were more strongly associated with teaching scores than was the total screening score (*r*s ranging from 0.04 to 0.23).

Table 6 reports the results of hierarchical multiple regression analyses testing how SJTs incrementally predicted interview day scores after accounting for the screening scores. At Step 1, total screening score made a statistically significant contribution to the prediction of all three interview day scores, with β-weights ranging from 0.22 to 0.26. The addition of SJT scores in Step 2 added incremental predictive validity to the interview scores variance ($\Delta R^2 = 0.15$, $p < .01$) and for the sample teaching ($\Delta R^2 = 0.20$, $p < .01$), but not for the group case study ($\Delta R^2 = 0.02$, $p =$ ns). All three regression equations were significant: interview score, $R^2 = 0.20$, $F(2, 94) = 11.80$, $p < .01$; group case study, $R^2 = 0.09$, $F(2, 94) = 4.47$, $p = .01$; and sample teaching, $R^2 = 0.24$, $F(2, 94) = 14.63$, $p < .01$.

The β-weight for the SJT was higher than the β-weight for the screening score for two out of the three interview day activities: interview score ($\beta_{SJT} = 0.40$, $p < .01$; $\beta_{SCREENING} = 0.12$, $p =$ ns), and sample teaching ($\beta_{SJT} = 0.46$, $p < .01$; $\beta_{SCREENING} = 0.07$, $p =$ ns). However, the β-weight for screening score was a better predictor of the group case study score than was the SJT ($\beta_{SJT} = .15$, $p =$ ns; $\beta_{SCREENING} = 0.21$, $p < .05$).

5. Discussion

We reported the development and validation of an online, construct-informed SJT to screen applicants who applied for a large and selective teacher education program in the UK. Results from the study suggest that a screening SJT was a reliable and valid predictor of interview day tasks and could be useful to screen applicants for more intensive selection approaches. Three research questions were posed in the study. In response to the first question, analysis of the psychometric properties of the revised SJT revealed acceptable internal consistency, and significant positive associations with concurrent screening and future interview day activities. We used a construct-informed approach to build SJT content, but the single factor structure emerging from the analyses did not reflect the targeted non-cognitive attributes used to develop item content. This pattern of results is not uncommon in SJT research because each 'real-life' scenario, even when built to target a particular construct, reflects multiple constructs (Campion, Ployhart, & MacKenzie, 2014). Our CFA and MAP analyses gives us some confidence that the SJT is measuring an overall factor of situational judgment, but the results raise issues about what role the foundation attributes targeted in the test's construction play. As has been shown in previous research, SJTs show promising levels of predictive validity, but there is a lack of clarity about which

Table 2
Correlations between SJT, screening test scores, and interview day scores.

	1	2	3	4	5	6	7	8
1. SJT	–	.30**	.30**	.35**	.42**	.20	.48**	.46**
2. Understanding/motivation essay (screening)		–	.58**	.82**	.14	.24*	.20	.24*
3. Leadership essay (screening)			–	.88**	.29*	.14	.14	.22*
4. Total screening				–	.24*	.26*	.22*	.29**
5. ID interview					–	.46**	.40**	.73**
6. ID case study						–	.56**	.84**
7. ID sample teaching							–	.85**
8. Total ID								–

Note. SJT = Situational Judgment Test; ID = Interview Day; *$p < .01$, **$p < .001$.

Table 3
Correlations between SJT, total screening score, and interview scores.

	1	2	3	4	5	6	7
1. SJT	–	.35**	.26**	.35**	.39**	.33**	.30**
2. Total screening		–	.13	.28**	.10	.22*	.19
3. Interview (Understanding and motivation)			–	.50**	.48**	.35**	.50**
4. Interview (Leadership)				–	.64**	.43**	.58**
5. Interview (Planning and organization)					–	.46**	.62**
6. Interview (Problem solving)						–	.35**
7. Interview (Resilience)							–

Note. SJT = Situational Judgment Test; *$p < .01$, **$p < .001$.

Table 4
Correlations between SJT, total screening score, and case study scores.

	1	2	3	4	5	6	7
1. SJT	–	.35**	.19	.20*	.06	.16	.25*
2. Total screening		–	.24*	.24*	.14	.25*	.25*
3. Case study (Empathy)			–	.81**	.65**	.60**	.50**
4. Case study (Interaction)				–	.69**	.66**	.58**
5. Case study (Leadership)					–	.71**	.47**
6. Case study (Problem solving)						–	.51**
7. Case study (Self-evaluation)							–

Note. SJT = Situational Judgment Test; *$p < .01$, **$p < .001$.

personal characteristics the tests are measuring (e.g., McDaniel et al., 2016).

The answer to the second question—does the SJT provide incremental predictive validity beyond current screening methods—was answered affirmatively, with evidence from hierarchical multiple regression analysis showing that scores on the SJT predict applicant performance at the interview day. The links between SJT and interview and sample teaching demonstration were significant and positive; however, the SJT was less clearly associated with applicant scores on the group case study. It is likely that the individualized nature of the SJT (*What should you do?*) is less useful in predicting the group dynamics assessed in the group case study activity. The third question, pertaining to high- and low-scorers on the SJT, is pertinent to decision-making based on test scores, with results showing that applicants who fared poorly on the SJT also fared poorly on the multiple activities that took place during the interview day.

Although the scores from the SJT did not differ according to SES background, there were significant differences on the SJT favoring females, and these differences were not found in the other screening methods. Similar patterns of gender differences have been seen in other SJT research (e.g., Whetzel, McDaniel, & Nguyen, 2008); these patterns are a potential concern in a profession where recruiting and retaining males presents challenges for many education systems (e.g., Pollitt & Oldfield, 2017). Further investigation into the reasons behind female applicants' better performance on SJTs is worth further scrutiny, and ITE programs that use SJTs for screening and selection will want to consider the implications of these gender differences.

The overall aim of selection procedures is to make decisions about the probability of applicants' future success, but selection methods range in cost (including time costs) and how effective they are in predicting success. Recent work by Klassen & Kim, 2019 showed that the cost and predictive utility of teacher selection methods were statistically unrelated, with overall prediction of objectively measured (i.e., not self-reported) teacher effectiveness generally low. Assessing applicants' non-cognitive attributes in a systematic, cost-effective, objective, and efficient way during a selection process presents real challenges for ITE programs, but using state-of-the-art, evidence-supported selection methods increases the likelihood of making better-informed and evidence-

Table 5
Correlations between SJT, total screening score, and sample teaching scores.

	1	2	3	4	5	6	7
1. SJT	–	.35**	.45**	.40**	.42**	.37**	.43**
2. Total screening		–	.23*	.21*	.22*	.21*	.04
3. Sample teaching (Empathy)			–	.74**	.61**	.77**	.60**
4. Sample teaching (Interaction)				–	.78**	.84**	.61**
5. Sample teaching (Planning and organization)					–	.68**	.57**
6. Sample teaching (Resilience)						–	.61**
7. Sample teaching (Self-evaluation)							–

Note. SJT = Situational Judgment Test; *$p < .01$, **$p < .001$.

Table 6
Hierarchical multiple regression analyses predicting interview day performance in interview, case study, and sample teaching from screening scores and SJTs.

Predictor	Assessment Center Activities					
	Interview		Case study		Sample teaching	
	ΔR^2	β	ΔR^2	β	ΔR^2	β
Step 1						
Total screening	.06*	.24*	.07*	.26*	.05*	.22*
Step 2						
Total screening		.12		.21*		.07
SJT	.15**	.40**	.02	.15	.20**	.46**
Total R^2	.20**		.09		.24**	
n	97		97		97	

Note. SJT = Situational Judgment Test; *$p < .01$, **$p < .001$.

supported selection decisions (Lievens & Sackett, 2017).

One important question about ITE selection is whether assessing non-cognitive attributes that may change over time are worth including in a selection process. There is little research in education on this topic, but evidence from other disciplines suggests that the relationship between the attributes measured at selection and targeted outcomes may evolve. Blair and colleagues (Blair, Hoffman, & Ladd, 2016) showed in a business setting that SJTs and general mental ability both significantly predicted work success one year after initial assessment, but performance in assessment centers did not. However, six years after the initial assessment, the contribution of general mental ability dissipated, SJTs continued to be related to work success (but to a lesser extent), and scores from assessment centers increased in their association with success. Similar results were found in selection into medicine and dentistry. Buyse and Lievens (2011) found that the predictive validity of cognitive ability measured at selection dropped through the five years of dental training, whereas the predictive validity of SJTs designed to assess interpersonal skills increased from negligible in Year 1 to positive and significant in Year 5. Similarly, Lievens and Sackett (2012) showed that an interpersonal SJT administered at selection into medical education grew in importance up to nine years into professional practice. Non-cognitive attributes may be nurtured during professional training, but the starting point—the core attributes measured at selection—appear to play an important role in future professional competence. For ITE programs, ignoring the evaluation of applicants' non-cognitive attributes at the point of selection, or using methods that lack an evidence base, may prove costly.

5.1. Limitations

The study does not address the longer-term predictive utility of the test, and further work is needed to connect SJTs and longer-term teaching outcomes during professional practice. Little is known about how methods used at the point of selection are related to later teaching outcomes (Goldhaber, Grout, & Huntington-Klein, 2014), and we do not yet know if the decisions made at selection are related to teaching effectiveness during professional practice. A recent meta-analysis suggests a weak relationship (Klassen & Kim, 2019), but the selection methods used in most of the studies in the meta-analysis were not reflective of recent advances in selection research.

We know that SJTs used for selection in professional fields tend to be positively related to professional outcomes (e.g., Lievens & Sackett, 2012), but what is less clear is which underlying constructs are contributing to the prediction of these outcomes. Our factor analyses showed that the data did not separate cleanly into the targeted non-cognitive attributes on which the SJT was built, but instead reflected a general judgment domain. This lack of clear factor structure is common in SJT research (e.g., Fröhlich et al., 2017), and points to the difficulty of separating out 'clean' factors when using complex, real-life scenarios as test stimuli. A next step in the development of SJTs for teacher selection may be to build tools that focus on single, well-defined constructs, such as integrity, emotional intelligence, and conscientiousness (see Libbrecht & Lievens, 2012 for a review). However, the hallmark of SJTs is their real-world relevance, and one advantage to using contextualized situations (i.e., a challenging classroom scenario) to evaluate judgment is that how people enact their personality, beliefs, and motivations is dependent on contextual factors (e.g., Chen, Fan, Zheng, & Hack, 2016). Finally, the study was conducted in one UK setting, and although the program is located in a large metropolitan area, further work on the cross-cultural application of SJTs is worth pursuing. Recent work exploring the universality of the non-cognitive attributes of effective teachers shows considerable overlap across settings, albeit with an overlay of culture-specific features (e.g., Klassen et al., 2018). Our program of research acknowledges the importance of cultural factors in developing teacher selection methods, and our current work is focused on developing SJTs and other selection methods in a range of non-UK and non-English speaking settings.

Acknowledgement

Funding for this article was received from the European Research Council Consolidator grant SELECTION 647234.

References

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

Bardach, L., & Klassen, R. M. (2019). *Smart teachers, successful students? A systematic review of the literature on teachers' cognitive abilities and teacher effectiveness.* .. https://psyarxiv.com/nt7v9

Bastian, K. C., McCord, D. M., Marks, J. T., & Carpenter, D. (2017). A temperament for teaching? Associations between personality traits and beginning teacher performance and retention. *AERA Open, 3*, 1–17.

Bergman, M. E., Donovan, M. A., Drasgow, F., 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.

Bieri, C., & Schuler, P. (2011). Cross-curricular competencies of student teachers: A selection model based on assessment centre admission tests and study success after the first year of teacher training. *Assessment & Evaluation in Higher Education, 36*, 399–415.

Blair, C. A., Hoffman, B. J., & Ladd, R. T. (2016). Assessment centers vs situational judgment tests: Longitudinal predictors of success. *The Leadership & Organization Development Journal, 37*, 899–911.

Buyse, T., & Lievens, F. (2011). Situational judgment tests as a new tool for dental student selection. *Journal of Dental Education, 75*, 743–749.

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., & Childs, R. (2011). Teacher education admission criteria as measure of

preparedness for teaching. *Canadian Journal of Education, 34*, 3–20.
Chang, M. L. (2013). Toward a theoretical model to understand teacher emotions and teacher burnout in the context of student misbehavior: Appraisal, regulation and coping. *Motivation and Emotion, 37*, 799–817.
Chen, L., Fan, J., Zheng, L., & Hack, E. (2016). Clearly defined constructs and specific situations are the currency of SJTs. *Industrial and Organizational Psychology, 9*, 34–38.
Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *The American Economic Review, 104*, 2593–2632.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
Cook, M. (2009). *Personnel selection: Adding value through people* (5th ed.). Chichester, UK: Wiley-Blackwell.
Davies, P., Connolly, M., Nelson, J., Hulme, M., Kirkman, J., & Greenway, C. (2016). 'Letting the right one in': Provider contexts for recruitment to initial teacher education in the United Kingdom. *Teaching and Teacher Education, 60*, 291–302.
Davison, H. K., & Burke, M. J. (2000). Sex discrimination in simulated employment contexts: A meta-analytic investigation. *Journal of Vocational Behavior, 56*, 225–248.
De Leng, W. E., Stegers-Jager, K. M., Husbands, A., Dowell, J. S., Born, M. P., & Themmen, A. P. N. (2017). Scoring method of a situational judgment test: Influence on internal consistency reliability, adverse impact and correlation with personality? *Advances in Health Sciences Education: Theory and Practice, 22*, 243–265.
Department for Education. (2018). *Schools, pupils and their characteristics: January 2018*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/719226/Schools_Pupils_and_their_Characteristics_2018_Main_Text.pdf.
Donaldson, G. (2011). *Teaching: Scotland's future. Report of a review of teacher education in Scotland*. Edinburgh: The Scottish Government.
ETS Praxis. (2018). Retrieved from: https://www.ets.org/praxis.
Fröhlich, M., Kahmann, J., & Kadmon, M. (2017). Development and psychometric examination of a German video-based situational judgment test for social competencies in medical school applicants. *International Journal of Selection and Assessment, 25*, 94–110.
Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457–472.
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*. Seattle, WA: University of Washington.
Greenberg, J., Walsh, K., & McKee, A. (2015). *Teacher prep review: A review of the nation's teacher preparation programs*. National Council on Teacher Quality. Retrieved from http://www.nctq.org/dmsView/Teacher_Prep_Review_2014_Report.
Guenole, N., Chernyshenko, O. S., & Weekly, J. (2017). On designing construct driven situational judgment tests: Somer preliminary recommendations. *International Journal of Testing, 17*, 234–252.
Hanushek, E. A., & Rivkin, S. G. (2012). The distribution of teacher quality and implications for policy. *Annual Review of Economics, 4*, 131–157.
Harris, D. N., & Rutledge, S. A. (2010). Models and predictors of teacher effectiveness: A review of the literature with lessons from (and for) other occupations. *Teachers College Record, 112*, 914–960.
Ilie, S., Sutherland, A., & Vignoles, A. (2017). Revisiting free school meal eligibility as a proxy for pupil socio-economic deprivation. *British Educational Research Journal, 43*, 253–274.
Ingvarson, L., & Rowley, G. (2017). Quality assurance in teacher education and outcomes: A study of 17 countries. *Educational Researcher, 46*, 177–193.
Kennedy, M. M., Ahn, S., & Choi, J. (2008). The value added by teacher education. In M. Cochran-Smith, S. Feiman-Nemser, D. J. McIntyre, & K. E. Demers (Eds.), *Handbook of research on teacher education* (3rd ed., pp. 1249–1273). New York, NY: Routledge.
Kim, L. E., Dar-Nimrod, I., & MacCann, C. (2018). Teacher personality and teacher effectiveness in secondary school: Personality predicts teacher support and student self-efficacy but not academic achievement. *Journal of Educational Psychology, 110*, 309–323.
Kim, L. E., & MacCann, C. (2018). Instructor personality matters for studentevaluations: Evidence from two subject areas at university. *British Journal of Educational Psychology, 88*, 584–605.
Klassen, R. M., & Durksen, T. L. (2015). Recent advances in research on teacher motivation and emotions. In C. Rubie-Davies (Ed.), *The social psychology of the classroom international handbook* (pp. 339–349). New York: Routledge, Taylor/Francis.
Klassen, R. M., Durksen, T. L., Al Hashmi, W., Kim, L. E., Longden, K., Metsäpelto, R.-L., Poikkeus, A. M., & Györi, J. (2018). Cultural context and teacher characteristics: Exploring the non-cognitive attributes of prospective teachers in four countries. Teaching and Teacher Education, 72, 64-74.
Klassen, R. M., Durksen, T. L., Kim, L. E., Patterson, F., Rowett, E., Warwick, J., & Wolpert, M. A. (2017). Developing a proof-of-concept selection test for entry into primary teacher education programs. *International Journal of Assessment Tools in Education, 4*, 96–114.
Klassen, R. M., & Kim, L. E. (2017). Assessing critical attributes of prospective teachers: Implications for selection into initial teacher education programs. In D. W. Putwain, & K. Smart (Eds.), *British Journal of Educational Psychology Monograph Series II: Psychological Aspects of Education* (pp. 5–22). Oxford: Wiley.
Klassen, R. M., & Kim, L. E. (2019). Selecting teachers and prospective teachers: A meta-analysis. *Educational Research Review, 26*, 32–51.
Klassen, R. M., & Tze, V. M. C. (2014). Teachers' self-efficacy, personality, and teaching effectiveness: A meta-analysis. *Educational Research Review, 12*, 59–76.
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.
Libbrecht, N., & Lievens, F. (2012). Validity evidence for the situational judgment test paradigm in emotional intelligence measurement. *International Journal of Psychology, 47*, 438–447.
Lievens, F. (2017). Construct-driven SJTs: Toward an agenda for future research. *International Journal of Testing, 17*, 269–276.
Lievens, F., & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology, 9*, 3–22.
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 success and job performance. *Journal of Applied Psychology, 97*, 460–468.
Lievens, F., & Sackett, P. R. (2017). The effects of predictor method factors on selection outcomes: A modular approach to personnel selection procedures. *Journal of Applied Psychology, 102*, 43–66.
Mason, R. W., & Schroeder, M. P. (2014). The predictive validity of teacher candidate letters of reference. *Journal of Education and Learning, 3*, 67–75.
Maulana, R., Helms-Lorenz, M., & van de Grift, W. (2015). A longitudinal study of induction on the acceleration of growth in teaching quality of beginning teachers through the eyes of their students. *Teaching and Teacher Education, 51*, 225–245.
McDaniel, M. A., List, S. K., & Kepes, S. (2016). The "hot mess" of situational judgment test construct validity and other issues. *Industrial and Organizational Psychology, 9*, 47–51.
McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730–740.
Melchers, K. G., & Kleinmann, M. (2016). Why situational judgment is a missing component in the theory of SJTs. *Industrial and Organizational Psychology, 9*, 29–34.
Metzger, S. A., & Wu, M.-J. (2008). Commercial teacher selection instruments: The validity of selecting teachers through beliefs, attitudes, and values. *Review of Educational Research, 78*, 921–940. https://doi.org/10.3102/0034654308323035.
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*, 749–761.
Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*, 313–335. https://doi.org/10.1037/a0026802.
Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., & Cleland, J. (2016). How effective are selection methods in medical education? A systematic review. *Medical Education, 50*, 36–60.
Pollitt, K., & Oldfield, J. (2017). Overcoming the odds: Exploring barriers and motivations for male trainee primary teachers. *Teaching and Teacher Education, 62*, 30–36.
Rimm-Kaufman, S. E., & Hamre, B. K. (2010). The role of psychological and developmental science in efforts to improve teacher quality. *Teachers College Record, 112*, 2988–3023.
Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy, 6*, 43–74.
Rockstuhl, T., Ang, S., Ng, K.-Y., Lievens, F., & Van Dyne, L. (2015). Putting judging situations into situational judgment tests: Evidence from intercultural multimedia SJTs. *Journal of Applied Psychology, 100*, 464–480.
van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Van Aken, M. A. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development, 85*(3), 842–860.
Schubert, S., Ortwein, H., Dumitsch, A., Schwantes, U., Wilhelm, O., & Kiessling, C. (2008). A situational judgement test of professional behaviour: Development development and validation. *Medical Teacher, 30*, 528–533.
Schumaker, G., Grigsby, B., & Vesey, W. (2013). Determining effective teaching behaviors through the hiring process. *International Journal of Educational Management, 29*, 139–155.
Seaton, F. S. (2017). Empowering teachers to implement a growth mindset. *Educational Psychology in Practice, 34*, 41–57.
Sutton, R. E. (2004). Emotional regulation goals and strategies of teachers. *Social Psychology of Education, 7*, 379–398.
Taxer, J. L., & Gross, J. J. (2018). Emotion regulation in teachers: The "why" and "how". *Teaching and Teacher Education, 74*, 180–189.
Taylor, N., Mehra, S., Elley, K., Patterson, F., & Cousans, F. (2016). The value of situational judgement tests for assessing non-academic attributes in dental selection. *British Dental Journal, 220*, 565–566.
Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. Goffin, & E. Helmes

(Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 41–71). Boston, MA: Kluwer.

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.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73*, 89–122.

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. In J. A. Weekley, & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement and application* (pp. 157–182). Mahwah, NJ: Lawrence Erlbaum Associates.

Weng, Q., Yang, H., Lievens, F., & McDaniel, M. A. (2018). Optimizing the validity of situational judgment tests: The importance of scoring methods. *Journal of Vocational Behavior, 104*, 199–209.

Whetzel, D. L., & McDaniel, M. A. (2016). Are situational judgment tests better assessments of personality than traditional personality tests in high-stakes testing? In U. Kumar (Ed.), *The Wiley handbook of personality assessment* (pp. 205–214). Chichester, UK: Wiley & Sons.

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance, 21*, 291–309.

Xu, Z., Özek, U., & Hansen, M. (2015). *Teacher performance trajectories in high-and lower-poverty schools. Educational Evaluation and Policy Analysis.* Retrieved from http://epa.sagepub.com/content/37/4/458.short.