


The selection gap in teacher education: Adverse effects of ethnicity, gender, and socio-economic status on situational judgement test performance

Lisa Bardach*¹ , Jade V. Rushby² and Robert M. Klassen²

¹University of Tübingen, Hector Research Institute of Education Sciences and Psychology, Germany

²Department of Education, University of York, UK

Background. Situational judgement tests (SJTs) measure non-cognitive attributes and have recently drawn attention as a selection method for initial teacher education programmes. To date, very little is known about adverse impact in teacher selection SJT performance.

Aims. This study aimed to shed light on adverse effects of gender, ethnicity, and socio-economic status (SES) on SJT scores, by exploring both main effects and interactions, and considering both overall SJT performance and separate SJT domain scores (mindset, emotion regulation, and conscientiousness).

Sample. A total of 2,808 prospective teachers from the United Kingdom completed the SJTs as part of the initial stage of selection into a teacher education programme.

Methods. In addition to SJT scores, the variables gender (female vs. male), ethnicity (majority group vs. minority group), and home SES background (higher SES status vs. lower SES status) were used in the analyses. Regression models and moderated regression models were employed.

Results and conclusions. Results from the regression models revealed that gender effects (females scoring higher than males) were restricted to emotion regulation, while ethnicity effects (ethnic majority group members scoring higher than ethnic minority group members) emerged for SJT overall scores and all three domains. Moderated regression modelling results furthermore showed significant interactions (gender and ethnicity) for SJT overall scores and two domains. Considering the importance of reducing subgroup differences in selection test scores to ensure equal access to teacher education, this study's findings are a critical contribution. The partially differentiated results for overall vs. domain-specific scores point towards the promise of applying a domain-level perspective in research on teacher selection SJTs.

Diversifying the teacher workforce has long been a concern of educational policy (e.g., Kirby, Berands, & Naftel, 1999). However, limited progress has been made to reach this goal (e.g., Albert Shanker Institute, 2015; OECD, 2016), as indicated by the relative scarcity of

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

*Correspondence should be addressed to Lisa Bardach, University of Tübingen, Hector Research Institute of Education Sciences and Psychology, Walter-Simon-Straße 12, 72072 Tübingen, Germany (email: lisa.bardach@uni-tuebingen.de).

minority group teachers (e.g., Nguyen & Redding, 2018) or male teachers in areas such as primary education (OECD, 2016). Focusing on the selection methods used by initial teacher education (ITE) providers, and exploring and eventually overcoming the potential adverse impact of selection tests can be seen as one starting point to widening participation. Adverse impact in selection practices occurs if the selection rate for a designated minority group is lower than that for the majority group, leading to systematic disadvantages for minority group members in the selection process (e.g., Ng & Sears, 2010).

Over the last few years, situational judgement tests (SJTs) have increasingly been used to inform decisions for personnel selection and for selection into different degree programmes (e.g., for medical school applicants, Fröhlich, Kahmann, & Kadmon, 2017; Lievens, 2013) and they have, more recently, successfully been applied in research on teacher selection (Klassen et al., 2017; Klassen, Durksen, Rowett, & Patterson, 2014; Klassen, Kim, Rushby, & Bardach, 2020). In a SJT, applicants are presented with scenarios they are likely to encounter during employment in the field. Following a contextualized description of each scenario, several potential ways to respond to the situation are provided, and the applicant has to judge the effectiveness of each response (e.g., Oostrom, Born, Serlie, & van der Molen, 2010). A solid body of evidence has been amassed on the criterion-related validity of SJTs (e.g., Teng, Brannick, & Borman, 2019) and their incremental validity over-and-above cognitive ability and personality tests (e.g., Christian, Edwards, & Bradeley, 2010), providing empirical support for their widespread use in selection settings. Moreover, it has been shown that SJTs produce fewer subgroup differences than cognitive ability tests (e.g., Lievens, Peeters, & Schollaert, 2008; Whetzel & McDaniel, 2009). Still, the existing body of SJT literature also documents, for instance, ethnicity and gender effects, with members of ethnic majority groups typically outperforming those of minority groups and with females outperforming males (e.g., Lievens, Patterson, Corstjens, Martin, & Nicholson, 2016; for a meta-analysis see Whetzel, McDaniel, & Nguyen, 2008).

To date, potential subgroup differences in SJT performance have not yet been sufficiently addressed in the context of teacher education. Gaining a better understanding of subgroup differences in teacher selection SJTs is critical from both a practical and theoretical viewpoint. Practically, ITE providers are challenged to make well-informed decisions regarding which selection tests to use, and which ones to replace or to abandon. Hence, they need comprehensive information on the characteristics of selection tests such as SJTs, including information on potential adverse impacts, in order to weigh advantages and disadvantages before deciding. Theoretically, we lack knowledge on the functioning of SJTs in subgroups of teacher applicants and carrying over assumptions from other populations (e.g., medical education students) might be inappropriate as different programmes attract different students with different motivations, abilities, career intentions, and attributes. Hence, a teacher education-specific perspective on SJTs and subgroup effects is warranted, calling for studies on SJTs conducted in the context of teacher education.

The present work therefore investigated key issues surrounding adverse impact in terms of gender, ethnicity, and socio-economic status (SES) on teacher selection SJT performance. With the aim of advancing the current knowledge of subgroup differences in SJT scores and providing potentially useful information for selection practice, we investigated both overall SJT scores and analysed the role of constructs with more granularity by considering scores on separate SJT domains (conscientiousness, mindset, emotion regulation). In addition to exploring main effects of gender, SES, and ethnicity and to provide information that benefits both research and practice, we furthermore strove to

achieve a more profound understanding by gaining insights into how these individual difference variables interact in predicting prospective teachers' SJT performance.

Teacher selection

'Teacher quality' constitutes the single most important school variable affecting student achievement (OECD, 2005) and being taught by an (in) effective teacher has long-term implications beyond students' school careers. For example, a study by Chetty et al. (2014) showed that students exposed to more effective teachers were more likely to attend college and earn higher salaries. To date, widespread consensus among researchers, economists, educators, and policymakers has been reached that improving the teacher workforce should lead to improved educational outcomes (e.g., Burroughs et al., 2019; Hanushek & Rivkin, 2012). Thereby, the selection of prospective teachers has been identified as a promising strategy to raise teacher quality (Klassen et al., 2014).

In the United Kingdom, where the data from this study were collected, applicants who are interested in training to become a teacher typically undergo a selection process consisting of three steps: First, they have to pass screening checks to ensure successful completion of appropriate qualifications, which is usually a first degree in a teachable subject area. Second, they need to successfully complete national-level literacy and numeracy tests (although this has recently been eliminated). Third, they have to participate in a face-to-face interview or assessment centre that includes a range of activities. The push to improve the quality of the teacher workforce (e.g., UK House of Commons Education Committee, 2012), however, has led to a growing interest in the development of new teacher selection methods, such as SJTs, to complement existing selection methods (e.g., Klassen et al., 2017; Klassen & Kim, 2017).

Using SJTs to assess prospective teachers' non-cognitive attributes

Situational judgement tests have been introduced to selection for teacher education as a way to improve the measurement of non-cognitive attributes, such as motivation and personality, at the point of selection into initial teacher training (see Klassen et al., 2014, 2018; Klassen & Kim, 2017 for an overview). A valid assessment of prospective teachers' non-cognitive attributes is critical, as non-cognitive attributes have relatively consistently been linked to teaching performance (e.g., Kim, Jörg, & Klassen, 2019; Klassen & Tze, 2014). Moreover, non-cognitive attributes might be at least as important as cognitive attributes, given that the latter show, at best, weak relations to teaching performance (for research syntheses see Aloe & Becker, 2009; Bardach & Klassen, 2020; D'Agostino & Powers, 2009). In contrast to conventional self-report questionnaires, SJTs hold the advantage of being less susceptible to socially desirable responses and faking (e.g., Nguyen, Bidermann, & McDaniel, 2005; Olaru et al., 2019), because they more indirectly and implicitly assess applicants' judgements of (in)appropriate responses (Johnson & Saboe, 2011; Motowidlo & Beier, 2010).

While the teacher selection SJTs cover several domains of non-cognitive attributes (see e.g., Klassen et al., 2014, 2017, 2020), the present study focuses on the domains of conscientiousness, mindset, and emotion regulation. These three domains were identified through an extensive literature review and discussions with ITE staff regarding critical characteristics of effective teachers (Klassen et al., 2020). The decision to include the personality trait conscientiousness was based on empirical evidence indicating that teachers scoring higher on conscientiousness tend to perform better in the classroom

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*)

Figure 1. Example of a situational judgment test.

(e.g., Baier et al., 2018; Kim et al., 2019). Mindset was chosen as one of the target attributes because teachers' beliefs about the nature of learning and the plasticity of student abilities can impact on their instructional choices as well as students' performance and self-beliefs (e.g., Roose, Vantieghem, Vanderlinde, & Van Avermaet, 2019; Timmerman, Kuyper, & van der Werf, 2015; Zhu, Urhahne, & Rubie-Davies, 2018). Finally, in consideration that everyday school life is replete with situations requiring the regulation of emotions in order to achieve beneficial educational outcomes (Frenzel, Becker-Kurz, Pekrun, & Goetz, 2015; Olson et al., 2019), emotion regulation was considered as a further target attribute. These three core attributes informed the content of the SJT analysed in this study. Figure 1 shows an example of an SJT developed for teacher selection.

Previous studies using these SJT items to assess prospective teachers' non-cognitive attributes have, for example, demonstrated positive relations between the SJT and other selection measures (concurrent validity, e.g., Klassen et al., 2020), hence pointing towards the promise of expanding the current teacher selection landscape by including SJTs. Still, the use of SJTs in teacher education is a relatively new and emerging line of research. As subgroup differences represent a core concern for the implementation of any selection system (Whetzel & McDaniel, 2009), investigating the potential for adverse impact ranks high among the research priorities of studies on SJTs for teacher selection.

Subgroup differences in SJT scores

To what extent is SJT performance prone to subgroup differences? Existing studies exploring gender differences indicate that on average, female test-takers show consistently better performance than male test-takers on SJTs (e.g., Lievens et al., 2016; Whetzel et al., 2008). A possible explanation for this finding, confirmed in the meta-analysis of Whetzel et al. (2008), relates to the 'personality load' of an SJT; that is, the extent to which they correlate with personality measures. Specifically, the higher the association between an SJT and the personality traits of conscientiousness and agreeableness, the larger the gender gap in test performance, given that females tend to report higher levels of conscientiousness and agreeableness in comparison to males (e.g., Costa, Terracciano, &

McCrae, 2001; Vecchione, Alessandri, Barbaranelli, & Caprara, 2012). So far, only two studies with prospective teachers have gathered information on gender gaps in SJT performance. In line with the existing SJT literature, Klassen et al. (2020) reported gender differences in scores on text-based SJTs (females > males). In a further study comparing different SJT formats, gender effects have been found to be limited to the strictly text-based SJT format, whereas scores on video-based SJTs remained unaffected by applicants' gender (Bardach, Rushby, Kim, & Klassen, 2020). Applying a domain-perspective on the teacher selection SJTs, one might suspect that the presumably more strongly 'personality-loaded' SJT items designed to measure conscientiousness should be more susceptible to gender effects than SJT items targeting other domains.

In addition to gender bias, an adverse impact in terms of ethnicity has been documented for SJTs (e.g., Whetzel et al., 2008, here referring to 'race'); however, it should be stressed that the magnitude of adverse impact on minority groups of SJTs is usually lower than those reported for cognitive ability tests (e.g., Whetzel & McDaniel, 2009). The use of video-based SJTs has been discussed as a way to decrease ethnic group test score gaps (Chan & Schmitt, 1997). Video-based SJTs rely on video-scenes with actors or animated characters to replace text content (e.g., Fröhlich et al., 2017; Juster et al., 2019). It has in fact been demonstrated that minority group members were less disadvantaged when video-based SJTs instead of text-based SJTs were employed (Chan & Schmitt, 1997). However, in a recent study with prospective teachers addressing the adverse impact of SJTs, ethnicity effects occurred in all of the three investigated SJT conditions (video-based with text, video-based without text, and text-based) (Bardach et al., 2020). This study relied on an overall SJT score rather than examining separate domains, which might have clouded our understanding regarding the impact of SJT format on subgroup performance. For example, Roth, Bobko, and Buster (2013) divided SJTs on a construct-level and found that SJTs assigned to an 'interpersonal category' disadvantaged Black participants (vs. White participants) to a lesser extent. Given that all of the SJT items developed for teacher selection represent challenging social situations, however, other mechanisms may be relevant for the current study. Specifically, considering that personality differences between ethnic groups tend to be negligible (e.g., Folds, Duehr, & Ones, 2008; Ones & Anderson, 2002), it is plausible that smaller ethnicity effects might emerge for the domain of conscientiousness reflecting individual differences in the personality trait of conscientiousness, and stronger effects for mindset and emotion regulation.

Furthermore, it can be argued that in any (selection) test situation, socio-economic hardships place applicants in disadvantaged positions, as they may have had less access to education in the past, less support from home, or face financial barriers interfering with (higher) education pathways and career choices (e.g., Crosnoe & Muller, 2014; Griffin & Hu, 2015). It is thus not surprising that SJT scores have been found to be influenced by SES in prior research, even though the effects were considerably smaller in size compared to those observed for cognitive tests (e.g., Lievens et al., 2016). Until now, the effect of SES on SJT performance has not yet been explored in research on teacher selection, suggesting a need for empirical investigations to take up this issue.

Lastly, although prior research has provided vital insights into subgroup differences in SJT scores, the lion's share of research has focused on main effects on SJT performance. Subgroup memberships, however, may interact in a more complex manner than can be captured when only estimating main effects. For instance, does identifying as a male and as a member from an ethnic minority group put an applicant in greater risk of achieving a lower score on an SJT task? Answering this, and related questions, requires researchers to

shift their focus from a sole consideration of the main effects of subgroup variables such as gender, ethnicity, and SES, to also investigating interactions among these predictors (see e.g., Griffin & Hu, 2015).

The present study

The purpose of this study was to examine in more depth whether gender, ethnicity, and SES are related to prospective teachers' SJT performance, and whether exploring the interaction of these factors can further contribute to our understanding of potential subgroup differences. In a first set of analyses, we investigated main effects of gender, ethnicity, and SES on overall SJT scores: Do ethnicity, SES, and gender predict SJT performance (*Research Question 1*)? We assumed that ethnicity would have an effect on SJT performance, with members from majority groups scoring higher than members from minority groups (e.g., Whetzel et al., 2008, *Hypothesis 1*), and that SES would affect SJT scores, favouring applicants from a higher SES background (e.g., Lievens et al., 2016, *Hypothesis 2*). Moreover, we expected that gender would influence SJT performance, with females outperforming males (e.g., Klassen et al., 2020, *Hypothesis 3*), given that the majority of SJT items employed in this study were text-based (see section on Measures). As a next step, we explored interactions between the three individual difference variables in the prediction of SJT scores (*Research Question 2*). Whereas the main effects addressed in research question 1 only provide insights into the presence of subgroup differences across all different subgroups, studying interactions offers more differentiated insights (e.g., gender may only have an effect for those from a low(er) SES background etc.). In addition to the interactions between each pair of individual difference variables (gender and SES, gender and ethnicity, ethnicity and SES), we included a three-way interaction to provide comprehensive information. A significant three-way interaction would indicate that, for example, the interaction between gender and SES depends on the levels of the third variable ethnicity, in that gender may only have an effect for those from a low(er) SES background in the presence of ethnic minority group membership. Please note that we did not specify concrete hypotheses for the interactions, given that this is the first study in the teacher education context to investigate interactions.

Second, we revisited the effects of, and the interplay between, SES, gender, and ethnicity in predicting SJT performance, but relied on separate SJT domain scores (conscientiousness, emotion regulation, and mindset) to uncover potentially differentiated effects for these three target attributes. We thus asked: Do ethnicity, SES, and gender predict SJT performance in the domains of conscientiousness, emotion regulation, and mindset (*Research Question 3*, main effects across all different subgroups)? It was hypothesized that SES, gender, and ethnicity should be related to SJT performance, with advantages for ethnic majority group members (*Hypothesis 4*), applicants from a high(er) SES background (*Hypothesis 5*), and for females (*Hypothesis 6*). Furthermore, while it stands to reason that gender effects emerge for all three domains, they might be stronger for conscientiousness than for mindset and emotion regulation. Females report higher levels of conscientiousness than males (e.g., Costa et al., 2001; Vecchione et al., 2012), which might also be reflected in the scoring patterns for SJTs assessing conscientiousness (Whetzel et al., 2008). On the other hand, the advantage of ethnic majority group membership could probably be smaller for the more 'personality-loaded' conscientiousness SJT than for mindset and emotion regulation, as personality differences between ethnic groups appear to be negligible (e.g., Folds et al., 2008). Finally, given that this study was the first to explore the potential value of studying interactions among individual difference variables in predicting separate SJT

domains, we did not feel confident enough to formulate specific hypotheses regarding interactions and conducted exploratory analyses in that regard. We therefore asked: Do the interactions between gender and ethnicity, the interaction between gender and SES, the interaction between ethnicity and SES, and the three-way interaction predict SJT performance for the three domains (*Research Question 4*)?

Method

Sample and procedure

The sample of this study comprised of 2,808 prospective teachers (mean age = 26.83 years, $SD = 8.02$). The participants responded to the SJT as part of the initial stage of selection into a teacher education programme in the United Kingdom which prepares students to become primary and secondary education teachers in a range of subjects. All of the participants had successfully completed the eligibility check for teacher training in the United Kingdom (e.g., acceptable A-level examination results in relevant subjects and an undergraduate degree [at level 2:1 or better] in a relevant teaching subject, see Klassen et al., 2020 for a more detailed description) prior to completing the SJT. The SJT was a component of the next hurdle, the online application process, and the participants completed the SJT at their convenience on the device of their choice. As participants completed the SJT as part of the initial screening phase of the application process and had not previously been assessed by another selection measure apart from the eligibility check, range restriction concerns were minimized, representing a considerable strength to the current study. Typically, studies on subgroup differences are based on incumbent samples, which have gone through an extensive selection process prior to entering the organization or study programme. As such, effect sizes from studies on incumbent samples tend to be downwardly biased. Incumbents have been selected for their job or study programme either by the specific selection measure under investigation (direct range restriction) or other selection measures correlated with the selection measure of interest (indirect range restriction). Consequently, applicant samples are more likely to contain lower scoring applicants, whereas incumbent samples are more likely to consist of higher scoring applicants. When incumbent samples are analysed, direct or indirect range restriction can thus give rise to substantially underestimated subgroup difference estimates (Bobko & Roth, 2013; Herde, Lievens, Jackson, Shalfröoshan, & Roth, 2019; Roth, Le, Oh, Van Iddekinge, & Robbins, 2017).

In total, 55.9% of the participants identified as being White, 11.4% as Asian (e.g., Asian or Asian British – Pakistani), 7.4% as Black (e.g., Black or Black British – African), 2.9% as multiple ethnic groups, 2.8% as other ethnic groups (e.g., other Asian background), and 20.6 chose the option ‘prefer not to say’ or did not respond to the question asking them to indicate their ethnicity. Moreover, 54.4% identified as female, 30.3% as male, 0.5% as non-binary, and 14.9% chose the option ‘prefer not to say’ or did not respond to the question. Finally, with regard to SES status, 11.8% of the applicants had received free school meals (FSM), 4.7% had received education maintenance allowance (EMA), 4.7% reported receiving both, and 52.6% had not previously been in receipt of FSM or EMA (26.1% chose not to specify). FSME refers to when students were in compulsory education (primary or secondary school). Students were eligible for FSM if their parents received benefits (e.g., Universal Credit, Income Support) or earned under a certain amount per year. Educational Maintenance Allowance was for 16–19 years old students in further education (e.g., college or sixth form completing A-levels, vocational subjects).

All stages of the research were reviewed and approved by the authors' university ethics review board and by the selection and recruitment team at the teacher education provider. The authors of the present article are neither formally affiliated with the teacher education provider in question, nor were they involved in making selection decisions. The data for this study were gathered as part of the extensive pilot testing of the SJT and the SJT was not used for selection decisions.

Measures

SJTs

The SJT analysed in this study comprised 11 items; four items evaluated the target domains of mindset and emotion management, respectively, and the remaining three SJT items measured conscientiousness. The majority of the SJT items relied on a text-based format, but one video-based SJT was included for each domain (resulting in a total of eight text-based and three video-based SJT items). Considering that previous research has shown that SJT presentation format can affect the presence of gender effects in SJT performance (Bardach et al., 2020), it was important to ensure that the number of video-based scenarios did not differ between domains.

Each scenario had four response options. Accordingly, applicants were asked to rate the appropriateness of each of the options, from (1) appropriate to (4) inappropriate, in consideration of what a beginning teacher should do in the circumstances described in the scenario. The scoring key was developed using a hybrid approach, which combines two independently generated keys (see Bergman, Drasgow, Donovan, Henning, & Juraska, 2006). As such, concordance panels with subject matter experts (SMEs) in the field determined the initial scoring key. The concordance panel included 26 teacher educators and early career teachers (77% female; 19% ethnic minority; mean age of 31.2 years) with a wide range of teaching subjects represented.

As a next step, revisions of the scoring key were made based upon the level of expert agreement, item quality, and the scoring patterns of the top ten per cent of applicants. The scoring followed the scoring system described by Patterson, Ashworth, and Good (2013), where points are allocated based on the extent to which participants' responses align with the established scoring key. For instance, if an applicant's response was in direct alignment with the scoring key, they were allocated three points, if their answer was one position away, they were allocated two points, if their answer was two positions away, they were allocated one point, and no points were awarded for answers three positions away. Therefore, there were 12 points available for each scenario (4 response options \times 3 maximum points) equating to a total available score of 132 (11 scenarios \times 12 maximum points). The reliability coefficients (Cronbach's alpha) for the SJT were .59, and the reliability coefficients for the sub-domains were .44 for mindset, .32 for conscientiousness, and .43 for emotion regulation. Researchers used to work with 'classical' Likert scale type survey items (e.g., personality measures) might consider these coefficients as low; however, our reliability estimates are aligned with typical SJT reliability estimates (e.g., Campion, Ployhart, & MacKenzie, 2014; Lievens & Sackett, 2006). Because they are multidimensional, SJTs almost always manifest lower internal consistency reliability than do other constructs such as personality survey scales or cognitive ability measures (e.g., Ployhart & McKenzie, 2011).

Other measures

The participants responded to questions asking them to indicate their gender, ethnicity, and SES status. For this study, we created the following dummy-coded categories for each of the three variables: (1) gender (0 = female, 1 = male), (2) ethnicity (0 = majority group, that is, White background, 1 = minority group, that is, participants from all other backgrounds), (3) home SES background (0 = high(er) SES status, that is, those participants who indicated that they had neither received FSM nor EMA, 1 = low(er) SES status, that is, those participants who reported having been eligible for FSM, EMA, or both). We decided to use these broader categories instead of more fine-grained ones, as some categories were under-represented (e.g., only a very small number of participants indicated having received both free school meals and educational maintenance allowance). With regards to ethnicity, a recent study in the context of teacher selection (Bardach et al., 2020) used the same categories and we have mirrored this for the sake of consistency and comparability.

Statistical analyses

All analyses were performed using the statistical software Mplus (version 8.2; Muthén & Muthén, 1998–2010) and relied on the robust maximum likelihood estimator (MLR) implemented in Mplus. MLR statistically corrects standard errors and chi-square test statistics for the departures from normality, meaning that non-normal distribution of the dependent variable cannot bias the findings (Muthén, Muthén, & Asparouhov, 2016). We estimated two regression models (Models 1a and 1b) for composite SJT scores as outcomes and two moderated regression models (Models 2a and 2b) for SJT domains as outcomes. Model 1a and 2a included main effects of gender, ethnicity, and SES to investigate whether these variables affect SJT performance. Models 1b and 2b additionally included the interactions between each pair of individual difference variables to gain insights into whether effects differed depending on specific combinations of subgroup membership, as well as the three-way interaction. It should be noted that the effects of gender, ethnicity, and SES in Model 1b and 2b are conditional main effects and should be interpreted as such; for example, a negative effect of gender (gender: female = 0, male = 1; a negative effects means that males score lower than females) solely pertains for those who come from an ethnic majority background (ethnicity: majority group member = 0, minority group member = 1) and have a high(er) SES background (SES: higher SES background = 0, lower SES background = 1). Similarly, the effect of ethnicity, with ethnic minority group members (coded as 1) scoring lower than ethnic majority group members (coded as 0), refers to those who are females (coded as 0) and come from a higher SES background (coded as 0). We relied on manifest mean SJT scores in both models and the measures of gender, ethnicity, and SES consisted of single indicators. Figure 2 shows Model 1b and Model 2b.

We report unstandardized and standardized coefficients for the main effects and the interactions. The standardized coefficients can be interpreted according to Cohen's guidelines (Cohen, 1988), with values over .10, .30, and .50 reflecting small, moderate, and large effect sizes, respectively. All significance testing was performed at the .05 level. In our study, the amount of missing data on the item level ranged between 0% and 26.1%. Full information maximum likelihood estimation (FIML; Enders, 2010) was used to deal with missing data. Because all of the missing data were on predictor variables and because Mplus would automatically apply listwise deletion in this case, we mentioned the predictors' variances in the Mplus MODEL command. This brings the predictors into the

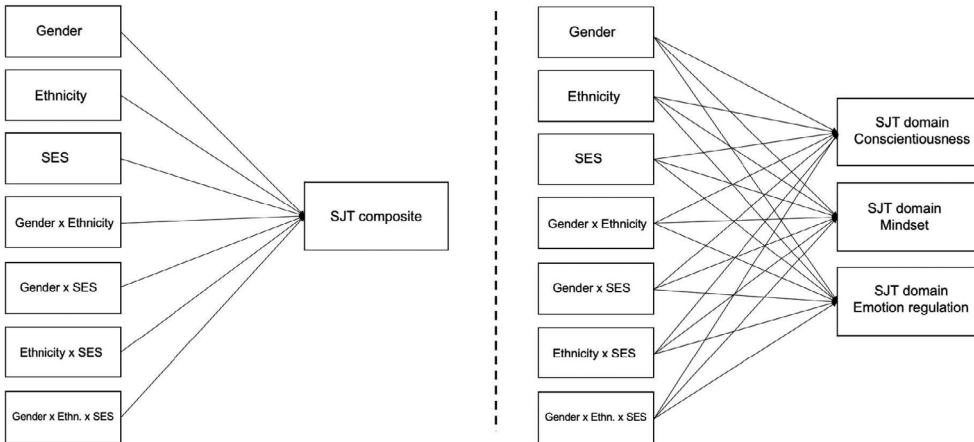


Figure 2. Graphical representation of Model 1b and Model 2b: Gender, ethnicity, SES, and interactions between all variables predicting SJT composite scores (Model 1b, left side) and gender, SES, and interactions between all variables predicting scores on the three SJT domains conscientiousness, mindset, and emotion regulation (Model 2b, right side).

models as dependent variables and allows them to be taken into account for FIML. However, even though this approach is widely used, some ambiguities remain as distributional assumptions are made about the predictors (e.g., multivariate normality). We therefore also ran all analyses using only complete cases. The same pattern of significant and non-significant results as in the main analyses emerged. These additional results can be obtained from the first author via request.

Results

Table 1 displays the descriptive statistics (mean, standard deviation) for SJT overall scores and SJT domain scores as well as bivariate correlations between all variables.

Table 1. Descriptive statistics of SJT scores and bivariate correlations between the variables investigated in the study

Variable	1.	2.	3.	4.	5.	6.	7.
1. SJT composite score							
2. SJT domain conscientiousness	.65						
3. SJT domain mindset	.71	.19					
4. SJT domain emotion regulation	.75	.24	.30				
5. Gender	-.09	-.08	-.03	-.06			
6. Ethnicity	-.29	-.14	-.22	-.24	-.00		
7. SES	.05	.02	.03	.04	-.01	.20	
M	108.89	28.71	40.43	40.28			
SD	6.29	2.77	3.00	3.14			

Notes. SJT = situational judgement test; SES = socio-economic Status; gender was coded as a dichotomous variable with 0 = female and 1 = male; ethnicity was coded as a dichotomous variable with 0 = majority and 1 = minority; SES was coded as a dichotomous variable with 0 = high(er) SES background and 1 = low(er) SES background; statistically significant correlation coefficient at $\alpha = .05$ are boldface.

Table 2 shows all standardized and unstandardized regression coefficients for Model 1a and Model 2a (models without interactions, Model 1a: SJT composite score, Model 2a: SJT domain scores), and Table 3 shows all standardized and unstandardized regression coefficients for Model 1b and Model 2b (models including interactions, Model 1b: SJT composite score, Model 2b: SJT domain scores).

The results of Model 1a (see Table 2) revealed that ethnicity significantly predicted SJT performance (SJT composite score), with members from ethnic majority groups performing better than members from minority groups (standardized $\hat{\beta} = -.293$, $p < .01$). Moreover, we found a significant effect of gender (standardized $\hat{\beta} = -.086$, $p < .01$) and no significant effect of SES (standardized $\hat{\beta} = .045$, $p > .05$) on SJT scores.

In Model 1b (see Table 3) additionally including the interactions, the conditional main effect of ethnicity (ethnicity = 0 vs. 1, gender = 0, SES = 0) was significant (standardized $\hat{\beta} = -.271$, $p < .01$). The conditional main effect of gender (gender = 0 vs. 1, ethnicity = 0, SES = 0) was not significant (standardized $\hat{\beta} = -.034$, $p > .05$). The conditional main effect of SES (SES = 0 vs. 1, gender = 0, ethnicity = 0) did not attain statistical significance (standardized $\hat{\beta} = .05$, $p > .05$). Furthermore, the interaction between ethnicity and gender proved significant (standardized $\hat{\beta} = -.130$, $p < .01$) and indicated that the gender effect changed if ethnicity was additionally considered. Specifically, a gender effect occurred such that males performed lower than females for those with ethnic minority group membership. The other two interactions (gender and SES, ethnicity and SES) were not significant (standardized $\hat{\beta} = -.011$, $p > .05$, and standardized $\hat{\beta} = .069$, $p > .05$, respectively). Similarly, the three-way interaction was not significant (standardized $\hat{\beta} = .046$, $p > .05$).

In Model 2a focusing on the three different domains (see Table 2), ethnicity significantly predicted SJT scores in the domain of conscientiousness (standardized

Table 2. Regression Models 1a and 2a: Unstandardized and standardized estimates of all effects

Effects	Unstandardized estimates (SE)	Standardized estimates (SE)
Model 1a: SJT composite score		
Gender → SJT scores	-1.119 (0.272)	-0.086 (0.021)
Ethnicity → SJT scores	-3.941 (0.309)	-0.293 (0.021)
SES → SJT scores	0.629 (0.322)	0.045 (0.023)
Model 2a: Domain conscientiousness		
Gender → SJT scores	-0.494 (0.120)	-0.085 (0.021)
Ethnicity → SJT scores	-0.868 (0.127)	-0.146 (0.021)
SES → SJT scores	0.149 (0.138)	0.024 (0.022)
Model 2a: Domain Mindset		
Gender → SJT scores	-0.194 (0.131)	-0.031 (0.021)
Ethnicity → SJT scores	-1.452 (0.148)	-0.227 (0.022)
SES → SJT scores	0.187 (0.157)	0.028 (0.024)
Model 2a: Domain emotion regulation		
Gender → SJT scores	-0.411 (0.135)	-0.063 (0.021)
Ethnicity → SJT scores	-1.650 (0.156)	-0.245 (0.022)
SES → SJT scores	0.299 (0.159)	0.043 (0.023)

Notes. Moderated regression results with $n = 2,808$. SE = standard error; gender was coded as a dichotomous variable with 0 = female and 1 = male; Ethnicity was coded as a dichotomous variable with 0 = majority and 1 = minority; SES was coded as a dichotomous variable with 0 = high(er) SES background and 1 = low(er) SES background; statistically significant results at $\alpha = .05$ are boldface.

Table 3. Moderated regression Models 1b and 2b: Unstandardized and standardized estimates of all effects

Effects	Unstandardized estimates (SE)	Standardized estimates (SE)
Model 1b: SJT composite score		
Gender → SJT scores	-0.443 (0.350)	-0.034 (0.027)
Ethnicity → SJT scores	-3.646 (0.491)	-0.271 (0.035)
SES → SJT scores	0.074 (0.477)	0.005 (0.034)
Gender × Ethnicity → SJT scores	-2.573 (0.951)	-0.130 (0.048)
Gender × SES → SJT scores	-0.244 (0.819)	-0.011 (0.039)
Ethnicity × SES → SJT scores	1.295 (0.868)	0.069 (0.046)
Ethnicity × Gender × SES → SJT scores	1.375 (1.563)	0.046 (0.052)
Model 2b: Domain conscientiousness		
Gender → SJT scores	-0.149 (0.165)	-0.026 (0.029)
Ethnicity → SJT scores	-0.648 (0.198)	-0.109 (0.033)
SES → SJT scores	0.233 (0.207)	0.038 (0.034)
Gender × Ethnicity → SJT scores	-0.983 (0.379)	-0.112 (0.043)
Gender × SES → SJT scores	0.011 (0.361)	-0.068 (0.040)
Ethnicity × SES → SJT scores	0.365 (0.299)	0.001 (0.043)
Ethnicity × Gender × SES → SJT scores	1.057 (0.657)	0.079 (0.049)
Model 2b: Domain mindset		
Gender → SJT scores	0.138 (0.171)	0.022 (0.027)
Ethnicity → SJT scores	-1.285 (0.233)	-0.201 (0.036)
SES → SJT scores	-0.049 (0.243)	-0.007 (0.037)
Gender × Ethnicity → SJT scores	-1.200 (0.447)	-0.127 (0.047)
Gender × SES → SJT scores	-0.085 (0.404)	-0.008 (0.040)
Ethnicity × SES → SJT scores	0.581 (0.404)	0.065 (0.045)
Ethnicity × Gender × SES → SJT scores	0.463 (0.742)	0.032 (0.052)
Model 2b: Domain emotion regulation		
Gender → SJT scores	-0.424 (0.174)	-0.065 (0.027)
Ethnicity → SJT scores	-1.754 (0.260)	-0.261 (0.038)
SES → SJT scores	-0.123 (0.226)	-0.018 (0.033)
Gender × Ethnicity → SJT scores	-0.369 (0.461)	-0.037 (0.046)
Gender × SES → SJT scores	0.528 (0.418)	0.050 (0.039)
Ethnicity × SES → SJT scores	0.736 (0.437)	0.078 (0.046)
Ethnicity × Gender × SES → SJT scores	-0.206 (0.780)	-0.014 (0.052)

Notes. Moderated regression results with $n = 2,808$. SE = standard error; gender was coded as a dichotomous variable with 0 = female and 1 = male; ethnicity was coded as a dichotomous variable with 0 = majority and 1 = minority; SES was coded as a dichotomous variable with 0 = high(er) SES background and 1 = low(er) SES background; statistically significant results at $\alpha = .05$ are boldface.

$\hat{\beta} = -.146$, $p < .01$, with ethnic majority group members > ethnic minority group members). Gender also significantly predicted conscientiousness scores (standardized $\hat{\beta} = -.085$, $p < .01$, females > males). There was no significant effect of SES on conscientiousness scores (standardized $\hat{\beta} = .024$, $p > .05$). For mindset, the results showed ethnicity effects (standardized $\hat{\beta} = -.227$, $p < .01$, ethnic majority group members > ethnic minority group members), but no effects of gender and SES (standardized $\hat{\beta} = -.031$, $p > .05$, and standardized $\hat{\beta} = .028$, $p > .05$, respectively). Ethnicity (lower scores for ethnic minority group members) and gender (lower scores for males) significantly predicted emotion management SJT scores (standardized $\hat{\beta} = -.245$,

$p < .01$, and standardized $\hat{\beta} = -.063$, $p < .01$, respectively). SES did not significantly predict mindset scores (standardized $\hat{\beta} = .043$, $p > .05$).

In Model 2b (see Table 3) including the interactions, for conscientiousness, the conditional main effect of ethnicity was significant (standardized $\hat{\beta} = -.109$, $p < .01$), whereas the conditional main effects of gender (standardized $\hat{\beta} = -.026$, $p > .05$) and SES (standardized $\hat{\beta} = .038$, $p > .05$) were not significant. A significant interaction between gender and ethnicity emerged, with gender effects (males < females) increasing in the presence of ethnic minority group membership (standardized $\hat{\beta} = -.112$, $p < .01$). The interactions between gender and SES and ethnicity and SES were not significant (standardized $\hat{\beta} = -.068$, $p > .05$, and standardized $\hat{\beta} = .001$, $p > .05$, respectively). The three-way interaction did not significantly predict SJT conscientiousness scores (standardized $\hat{\beta} = .079$, $p > .05$). With regard to mindset SJT performance, we obtained a significant conditional main effect of ethnicity (standardized $\hat{\beta} = -.201$, $p < .01$), and no significant conditional main effects of gender and SES (standardized $\hat{\beta} = .022$, $p > .05$, and standardized $\hat{\beta} = -.007$, $p > .05$, respectively). As for conscientiousness, the findings indicated a significant interaction between gender and ethnicity (standardized $\hat{\beta} = -.127$, $p < .01$). The interactions between gender and SES ($\hat{\beta} = -.008$, $p > .05$), ethnicity and SES ($\hat{\beta} = .065$, $p > .05$), as well as the three-way interaction ($\hat{\beta} = .032$, $p > .05$) were not significant. Finally, for emotion regulation, the findings revealed a significant conditional main effect for ethnicity (standardized $\hat{\beta} = -.261$, $p < .01$), and gender (standardized $\hat{\beta} = -.065$, $p < .05$), but not for SES (standardized $\hat{\beta} = -.018$, $p > .05$). None of the interactions reached statistical significance (ethnicity and gender: standardized $\hat{\beta} = -.037$, $p > .05$, gender and SES: standardized $\hat{\beta} = .050$, $p > .05$, SES and ethnicity: standardized $\hat{\beta} = .078$, $p > .05$, three-way interaction: standardized $\hat{\beta} = -.014$, $p > .05$).

Discussion

Despite the widely acknowledged importance of reducing subgroup differences in selection test scores to ensure equal access to education and relatedly, equal employment opportunities (e.g., Griffin & Hu, 2015), little is known about subgroup differences in SJT performance in the context of teacher selection. The present work therefore addressed adverse impact in terms of gender, ethnicity, and SES, considering both main effects and interactions and relying on overall SJT scores as well as domain-specific scores. With regard to main effects on SJT overall scores (Model 1a, *Research Question 1*), we found that males experienced a significant adverse impact (*Hypothesis 3*). This result is in accordance with our hypothesis suggesting that females would outperform males, meta-analytic findings (Whetzel et al., 2008), and the study by Klassen et al. (2020) with a text-based teacher selection SJT. It does not align with the findings of Bardach et al. (2020) who did not find any gender differences in conditions with video-based SJT items as compared to a text-based SJT condition. Whether the presence of gender effects in our study might thus be due to the fact that only three video-based SJTs were included – in addition to text-based ones – needs to be further clarified in future studies. In light of the higher costs involved in creating video-based SJTs, a related and practically relevant research question then becomes ‘how many video SJTs are enough to avoid adverse impact in terms of gender?’. Potential spillover effects could also be worth investigating, as the ratio of text vs. video-based SJTs might be less important than the fact that there are a certain number of video SJTs included.

Furthermore, our results indicated the known effects of ethnicity on SJT scores (e.g., Whetzel et al., 2008), whereby members from ethnic majority groups outperform members from ethnic minority groups (*Hypothesis 1*). This finding reinforces serious concerns and has important implications for future research related to investigating and appropriately modifying existing SJT test content and formats to ensure that ethnic minority group members are not adversely impacted by the use of SJTs for teacher selection (e.g., Lievens et al., 2008; Roth et al., 2013). On a more general level, a further important avenue for future research would be to test whether SJTs measure the same competencies across different ethnic groups. On the other hand, this study did not confirm that applicants who had been socio-economically less advantaged scored lower on the SJT than those from high(er) SES backgrounds (*Hypothesis 2*, e.g., Lievens et al., 2016). This is a positive finding; however, as this is the first study shedding light on SES as a potential predictor of SJT performance in the teacher education context, further studies are needed to investigate the robustness of our results. Thereby, we envision these studies should employ a range of SES proxies, given that different results might be obtained depending on the measurement of SES (e.g., Festin, Thomas, Ekberg, & Kristenson, 2017).

Re-estimating the model including interactions (Model 1b, *Research Question 2*) revealed several noteworthy findings that change the interpretation of the main effect reported above. First, in this model, differentiated gender effects were shown (i.e., the negative conditional main effect of gender predicting SJT scores was not significant), whereas the effect of ethnicity on SJT performance and the effect of SES on SJT performance remained significant and non-significant, respectively. Second, analysing the interplay between the three individual difference variables in forecasting SJT scores indicated a significant interaction between gender and ethnicity. Hence, gender effects (males performing less well than females) increased and became significant in the presence of ethnic minority group membership. No significant effects were reported for the interaction between SES and gender, ethnicity and SES, and the three-way interaction. The significant interaction between ethnicity and gender found in this study, however, yields important new knowledge on subgroup differences in SJT scores and points towards a group at risk of performing poorly on the SJT. Prior studies solely investigated whether being male or identifying as ethnic minority group member affects SJT scores in isolation from each other. By contrast, our study shows that it is the combination that matters and that can be linked to lower overall SJT performance.

In conclusion, our results for the overall SJT scores are of high relevance, particularly if we consider that selection decisions usually draw on composite scores. Still, the reliance on overall instead of domain-specific scores potentially masks meaningful heterogeneities. Following calls for more construct-driven perspectives on SJTs (e.g., Roth et al., 2013; also see e.g. Lievens, 2017), we therefore also ran all analyses using scores on mindset, conscientiousness, and emotion regulation as outcomes.

The results for the separate domains from Model 2a (*Research Question 3*, main effects) revealed that significant small gender effects (females > males) were restricted to conscientiousness and emotion regulation, thus only partially supporting our assumptions that gender should be related to SJT scores in all domains (*Hypothesis 6*). Interestingly, and aligned with prior research indicating that females describe themselves as more conscientious than males (e.g., Costa et al., 2001; Vecchione et al., 2012), conscientiousness was one of the two dimensions significantly impacted by gender effects. We obtained an effect for SJTs focusing on emotion regulation too, which could probably be traced back to gender differences in the way emotions are managed, and emotion management skills and strategies (for a review on gender and emotion

management in the teacher domain, see e.g., Olson et al., 2019). Moreover, in accordance with the findings for SJT overall scores, low(er) SES background was not related to performance decrements in any of the domains, thus contradicting our hypothesis (*Hypothesis 5*). As hypothesized, however, ethnicity was significantly related to SJT scores in all three domains (majority group members > minority group members, *Hypothesis 4*). In addition, stronger ethnicity effects occurred for the less personality-focused domains of mindset and emotion regulation than for conscientiousness.

In Model 2b, which was set up to investigate interactions and their effects on SJT dimensions (*Research Question 4*), all conditional main effects for ethnicity and none of the conditional main effects for SES was significant. Moreover, we found a significant conditional main effect for gender only for the emotion regulation SJT. As was the case with overall SJT performance, neither the two interactions involving SES (gender and SES, ethnicity and SES) nor the three-way interaction turned out to be significant. Importantly, however, the interaction between gender and ethnicity attained statistical significance for conscientiousness and mindset, which mirrored the interaction found for SJT total scores. Significant gender effects in terms of increased disadvantages for males were coupled with belonging to an ethnic minority group. The reason why no such interaction emerged for emotion regulation cannot be confidently answered based on the current study and data. For example, it could be that for this specific group, some of the content covered in the conscientiousness and mindset may be more difficult to solve, irrelevant, or disconnected from their own prior educational experiences, thus leading to lower performance. It may also be that the male applicants with ethnic minority background in our sample approached the tasks of these two domains in a different way that is not necessarily 'wrong' but simply less aligned with the scoring key. Clearly, this and related issues remain to be more deeply explored in future research. Nonetheless, as a first study on interactions between three key individual difference variables (gender, ethnicity, SES), the findings for separate domains are particularly enlightening. Even though they share some commonalities with the result based on overall SJT scores (e.g., main effects and conditional main effects for ethnicity), they also indicate construct-specific patterns.

Limitations and future lines of research

A salient limitation of our study is that we focused on the measurement technique of SJTs and did not include data from other tests or other data sources than applicants (e.g., interview evaluation scores). Future research should therefore expand our work by focusing on a range of selection methods. Second, our study is inherently limited by its cross-sectional nature. Employing designs with repeated measurements could yield pivotal further insights, for instance on the temporal stability of subgroup differences in SJT scores or on relations between SJT performance at the day of selection into the programme and competence-related developmental trajectories over the course of teacher education, which might differ depending on subgroup membership. Third, it would be beneficial to use more fine-grained categorizations of individual difference variables. However, this was not possible in the present study due to the small number of participants from certain subgroups. Fourth, researchers used to work with 'classical' survey scales might point out that the reliability coefficients in our study are rather low, particularly for conscientiousness. Nevertheless, our reliability estimates are consistent with mean reliability estimates for the SJT format. SJTs typically demonstrate low internal consistency as they present multidimensional situations and response options (e.g., Campion et al., 2014; Lievens & Sackett, 2006).

Conclusions

In this paper, we presented the first study using SJTs that comprehensively investigated adverse impact in the context of teacher selection, focusing on gender, ethnicity, and SES. While our results converge with some previous research, for instance with regard to the effects of ethnicity (e.g., Whetzel et al., 2008; for SJTs in teacher selection research see Bardach et al., 2020), they also extend thinking about adverse impacts for SJT performance. In general, our results affirm the notion that exploring interactions in addition to main effects deepens our understanding of subgroup effects and holds important implications for selection practice. Moreover, we believe that the approach of focusing on SJT domains needs to be scrutinized in future teacher selection studies as the differentiated pattern of findings for domain-level vs. overall SJT scores obtained in our study imparts some confidence on the usefulness of studying separate domains.

Funding

This work is supported by the European Research Council (grant #647234 SELECTION).

Conflicts of interest

All authors declare no conflict of interest.

Author contribution

Lisa Bardach, Ph.D. (Conceptualization; Formal analysis; Methodology; Writing – original draft; Writing – review & editing) Jade Rushby (Writing – original draft; Writing – review & editing) Robert Klassen (Conceptualization; Funding acquisition; Project administration; Supervision; Writing – review & editing).

Data availability statement

The datasets analysed during the current study are available from the first author on reasonable request.

References

- Albert Shanker Institute (2015). *The state of teacher diversity in American education*. Washington, DC: Author. Retrieved from <http://www.shankerinstitute.org/resource/teacherdiversity>
- Aloe, A. M., & Becker, B. J. (2009). Teacher verbal ability and school outcomes: Where is the evidence? *Educational Researcher*, 38, 612–624. <https://doi.org/10.3102/0013189X09353939>.
- Baier, F., Decker, A.-T., Voss, T., Kleickmann, T., Klusmann, U., & Kunter, M. (2018). What makes a good teacher? The relative importance of mathematics teachers' cognitive ability, personality, knowledge, beliefs, and motivation for instructional quality. *British Journal of Educational Psychology*, 89, 767–786. <https://doi.org/10.1111/bjep.12256>.
- Bardach, L., & Klassen, R. M. (2020). Smart teachers, successful students? A systematic review of the literature on teachers' cognitive abilities and teacher effectiveness. *Educational Research Review*, 30, 100312. <https://doi.org/10.1016/j.edurev.2020.100312>.

- Bardach, L., Rushby, J. V., Kim, L. E., & Klassen, R. M. (2020). Using video-based situational judgment tests for teacher selection: A quasi-experiment exploring the relations between test format, subgroup differences, and applicant reactions. *European Journal of Work and Organizational Psychology*. Advance Online Publication. <https://doi.org/10.1080/1359432X.2020.1736619>.
- 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*(3), 223–235. <https://doi.org/10.1080/1359432X.2020.1736619>.
- Bobko, P., & Roth, P. L. (2013). Reviewing, categorizing, and analyzing the literature on Black-White mean differences for predictors of job performance: Verifying some perceptions and updating/correcting others. *Personnel Psychology*, *66*, 91–126. <https://doi.org/10.1111/peps.12007>
- Burroughs, N. A., Gardner, J., Lee, Y., Guo, S., Touitou, I., Jansen, K., & Schmidt, W. (2019). A review of the literature on teacher effectiveness and student outcomes. In N. A. Burroughs, J. Gardner, Y. Lee, S. Guo, I. Touitou, K. Jansen & W. Schmidt (Eds.), *Teaching for excellence and equity, IAE research for education* (Vol. 6, pp. 7–17). Cham, Switzerland: Springer Nature Switzerland.
- 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. <https://doi.org/10.1080/08959285.2014.929693>.
- Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology*, *82*, 143–159. <https://doi.org/10.1037/0021-9010.82.1.143>.
- Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, *104*, 2633–2679. <https://doi.org/10.1257/aer.104.9.2633>
- Christian, M. S., Edwards, B. D., & Bradeley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, *63*, 83–117. <https://doi.org/10.1111/j.1744-6570.2009.01163.x>.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Costa, Jr, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology*, *81*(2), 322–331. <https://doi.org/10.1037/0022-3514.81.2.322>.
- Crosnoe, R., & Muller, C. (2014). Family socioeconomic status, peers, and the path to college. *Social Problems*, *61*, 602–624. <https://doi.org/10.1525/sp.2014.12255>.
- D'Agostino, J. V., & Powers, S. J. (2009). Predicting teacher performance with test scores and grade point average: A meta-analysis. *American Educational Research Journal*, *46*, 146–182. <https://doi.org/10.3102/0002831208323280>.
- Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford.
- Festini, K., Thomas, K., Ekberg, J., & Kristenson, M. (2017). Choice of measure matters: A study of the relationship between socioeconomic status and psychosocial resources in a middle-aged normal population. *PLoS One*, *12*(8), e0178929. <https://doi.org/10.1371/journal.pone.0178929>.
- Foldes, H. J., Duehr, E. E., & Ones, D. S. (2008). Group differences in personality: Meta-analyses comparing five US racial groups. *Personnel Psychology*, *61*, 579–616. <https://doi.org/10.1111/j.1744-6570.2008.00123.x>.
- Frenzel, A. C., Becker-Kurz, B., Pekrun, R., & Goetz, T. (2015). Teaching this class drives me nuts! – Examining the person and context specificity of teacher emotions. *PLoS One*, *10*(6), e0129630. <https://doi.org/10.1371/journal.pone.0129630>.
- 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*(1), 94–110. <https://doi.org/10.1111/ijsa.12163>.

- Griffin, B., & Hu, W. (2015). The interaction of socio-economic status and gender in widening participation in medicine. *Medical Education*, *49*(1), 103–113. <https://doi.org/10.1111/medu.12480>.
- Hanushek, E. A., & Rivkin, S. G. (2012). The distribution of teacher quality and implications for policy. *Annual Review of Economics*, *4*, 131–157. <https://doi.org/10.1146/annurev-economic-s080511-111001>.
- Herde, C. N., Lievens, F., Jackson, D. J., Shalfrooshan, A., & Roth, P. L. (2019). Subgroup differences in situational judgment test scores: Evidence from large applicant samples. *International Journal of Selection and Assessment*, *28*, 45–54.
- Johnson, R., & Saboe, K. (2011). Measuring implicit traits in organizational research: Development of an indirect measure of employee implicit self-concept. *Organizational Research Methods*, *14*, 530–547. <https://doi.org/10.1177/1094428110363617>.
- Juster, F. R., Baum, R. C., Zou, C., Risucci, D., Ly, A., Reiter, H., . . . Dore, K. L. (2019). Addressing the diversity-validity dilemma using situational judgment tests. *Academic Medicine*, *94*, 1197–1203. <https://doi.org/10.1097/ACM.0000000000002769>.
- Kim, L., Jörg, V., & Klassen, R. M. (2019). A meta-analysis of the effects of teacher personality on teacher effectiveness and burnout. *Educational Psychology Review*, *31*, 163–195. <https://doi.org/10.1007/s10648-018-9458-2>.
- Kirby, S. N., Berands, M., & Naftel, S. (1999). Supply and demand of minority teachers in Texas: Problems and prospects. *Educational Evaluation and Policy Analysis*, *11*, 301–323. <https://doi.org/10.3102/01623737021001047>.
- Klassen, R. M., Durksen, T. L., Al Hashmi, W., Kim, L. E., Longden, K., Metsäpelto, R.-L., . . . Györi, J. G. (2018). National context and teacher characteristics: Exploring the critical non-cognitive attributes of novice teachers in four countries. *Teaching and Teacher Education*, *72*, 64–74. <https://doi.org/10.1016/j.tate.2018.03.001>.
- Klassen, R. M., Durksen, T., Kim, L., 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 Education*, *4*, 96–114. <https://doi.org/10.21449/ijate.275772>.
- Klassen, R., Durksen, T., 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*(2), 104–124. <https://doi.org/10.4471/ijep.2014.07>.
- 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*, (5–22). Oxford, UK: Wiley.
- Klassen, R. M., Kim, L. E., Rushby, J. V., & Bardach, L. (2020). Can we improve how we screen applicants for initial teacher education? *Teaching and Teacher Education*, *87*, 102949. <https://doi.org/10.1016/j.tate.2019.102949>.
- Klassen, R., & Tze, V. M. C. (2014). Teachers' self-efficacy, personality, and teaching effectiveness: A meta-analysis. *Educational Research Review*, *12*, 59–76. <https://doi.org/10.1016/j.edurev.2014.06.001>.
- Lievens, F. (2013). Adjusting medical school admission: Assessing interpersonal skills using situational judgement tests. *Medical Education*, *47*(2), 182–189. <https://doi.org/10.1111/medu.12089>.
- Lievens, F. (2017). Construct-driven SJTs: Toward an agenda for future research. *International Journal of Testing*, *17*, 269–276. <https://doi.org/10.1080/15305058.2017.1309857>.
- Lievens, F., Patterson, F., Corstjens, J., Martin, S., & Nicholson, S. (2016). Widening access in selection using situational judgement tests: Evidence from the UKCAT. *Medical Education*, *50*, 624–636. <https://doi.org/10.1111/medu.13060>.
- Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, *37*, 426–441. <https://doi.org/10.1108/00483480810877598>.

- Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology, 91*, 1181–1188. <https://doi.org/10.1037/0021-9010.91.5.1181>.
- 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*(2), 321–333. <https://doi.org/10.1037/a0017975>.
- Muthén, B. O., Muthén, L. K., & Asparouhov, T. (2016). *Regression and mediation analysis using Mplus*. Los Angeles, CA: Muthén & Muthén.
- Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Author.
- Ng, E. S., & Sears, G. J. (2010). The effect of adverse impact in selection practices on organizational diversity: A field study. *The International Journal of Human Resource Management, 21*, 1454–1471. <https://doi.org/10.1080/09585192.2010.488448>.
- Nguyen, N. T., Biderman, M. D., & McDaniel, M. A. (2005). Effects of response instructions on faking a situational judgment test. *International Journal of Selection and Assessment, 13*, 250–260. <https://doi.org/10.1111/j.1468-2389.2005.00322.x>.
- Nguyen, T. D., & Redding, C. (2018). Changes in the demographics, qualifications, and turnover of American STEM Teachers, 1988–2012. *AERA Open, 4*(3), 1–13. <https://doi.org/10.1177/2332858418802790>.
- OECD (2005). *Teachers matter: Attracting, developing and retaining effective teachers*. Paris, France: OECD Publishing. Retrieved from <http://www.oecd.org/education/school/attractingdevelopingandretainingeffectiveteachers-finalreportteachersmatter.html>
- OECD (2016). *Teachers, the learning environment and the organisation of schools*. Paris, France: Author. <https://doi.org/10.1787/0c41fec4-en>.
- Olaru, G., Burrus, J., MacCann, C., Zaromb, F. M., Wilhelm, O., & Roberts, R. D. (2019). Situational judgment tests as a method for measuring personality: Development and validity evidence for a test of dependability. *PLoS One, 14*(2), 1–19. <https://doi.org/10.1371/journal.pone.0211884>.
- Olson, R. E., McKenzie, J., Mills, K. A., Patulny, R., Bellocchi, A., & Caristo, F. (2019). Gendered emotion management and teacher outcomes in secondary school teaching: A review. *Teaching and Teacher Education, 80*, 128–144. <https://doi.org/10.1016/j.tate.2019.01.010>.
- Ones, D. S., & Anderson, N. (2002). Gender and ethnic group differences on personality scales in selection: Some British data. *Journal of Occupational and Organizational Psychology, 75*, 255–276. <https://doi.org/10.1348/096317902320369703>.
- Oostrom, J. K., Born, M. P., Serlie, A. W., & van der Molen, H. T. (2010). Webcam testing: Validation of an innovative open-ended multimedia test. *European Journal of Work and Organizational Psychology, 19*, 532–550. <https://doi.org/10.1080/13594320903000005>.
- Patterson, F., Ashworth, V., & Good, D. (2013). *Situational judgement tests: A guide for applicants to the UK Foundation Programme*. Medical Schools Council, 1–29.
- Ployhart, R. E., & MacKenzie, Jr, W. I. (2011). Situational judgment tests: A critical review and agenda for the future. In S. Zedeck (Ed.), *APA Handbooks in Psychology. APA handbook of industrial and organizational psychology, Vol. 2. Selecting and developing members for the organization* (pp. 237–252). Washington, DC: American Psychological Association.
- Roose, I., Vantieghe, W., Vanderlinde, R., & Van Avermaet, P. (2019). Beliefs as filters for comparing inclusive classroom situations. Connecting teachers' beliefs about teaching diverse learners to their noticing of inclusive classroom characteristics in video clips. *Contemporary Educational Psychology, 56*, 140–151. <https://doi.org/10.1016/j.cedpsych.2019.01.002>.
- Roth, P. L., Bobko, P., & Buster, M. A. (2013). Situational judgment tests: The influence and importance of applicant status and targeted constructs on estimates of Black-White subgroup differences. *Journal of Occupational and Organizational Psychology, 86*, 394–409. <https://doi.org/10.1111/joop.12013>.
- Roth, P. L., Le, H., Oh, I.-S., Van Iddekinge, C. H., & Robbins, S. B. (2017). Who r u?: On the (in) accuracy of incumbent-based estimates of range restriction in criterion-related and differential validity research. *Journal of Applied Psychology, 102*, 802–828. <https://doi.org/10.1037/apl0000193>.

- Teng, Y., Brannick, M. T., & Borman, W. C. (2019). Capturing resilience in context: Development and validation of a situational judgment test of resilience. *Human Performance, 33*, 74–103. <https://doi.org/10.1080/08959285.2019.1709069>.
- Timmermans, A. C., Kuyper, H., & van der Werf, G. (2015). Accurate, inaccurate, or biased teacher expectations: Do Dutch teachers differ in their expectations at the end of primary education? *British Journal of Educational Psychology, 85*, 459–478. <https://doi.org/10.1111/bjep.12087>.
- UK House of Commons Education Committee (2012). Great teachers: attracting, training and retaining the best. Government Response to the Committee's Ninth Report of Session 2010–12. House of Commons, London: The Stationery Office Limited. Retrieved from: <https://publications.parliament.uk/pa/cm201213/cmselect/cmeduc/524/524.pdf>
- Vecchione, M., Alessandri, G., Barbaranelli, C., & Caprara, G. (2012). Gender differences in the Big Five personality development: A longitudinal investigation from late adolescence to emerging adulthood. *Personality and Individual Differences, 53*, 740–746. <https://doi.org/10.1016/j.paid.2012.05.033>.
- Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review, 19*(3), 188–202. <https://doi.org/10.1016/j.HRMR.2009.03.007>.
- 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. <https://doi.org/10.1080/08959280802137820>.
- Zhu, M., Urhahne, D., & Rubie-Davies, C. M. (2018). The longitudinal effects of teacher judgement and different teacher treatment on students' academic outcomes. *Educational Psychology, 38*, 648–668. <https://doi.org/10.1080/01443410.2017.1412399>.

Received 2 August 2020; revised version received 14 December 2020