Teacher Selection Project Working Paper
31.07.19

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

Lisa Bardach, Jade V. Rushby and Robert M. Klassen
University of York, UK

This *Working Paper* reports in-progress work that is not yet peer reviewed. The purpose of the *Working Paper* is to stimulate discussion and to contribute to the advancement of knowledge. Please check our project website for news of recent publications. This work is supported by the European Research Council (grant #647234 SELECTION)

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

Lisa Bardach, Jade V. Rushby, and Robert M. Klassen
University of York

Abstract

Background: Situational judgment tests (SJTs) measure non-cognitive attributes and have recently drawn attention as a selection method for initial teacher education programs. 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 completed the SJTs as part of the initial stage of selection into a teacher education program.

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. Moderated regression models were employed.

Results and conclusions: 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. The results revealed significant interactions (gender and ethnicity; ethnicity and SES) 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.

Introduction

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 Black and other 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.

Over the last few years, situational judgment tests (SJTs) have increasingly been used to inform decisions for personnel selection and the selection into different degree programs (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, Durksen, Rowett, & Patterson, 2014; Klassen & Kim, 2017; Klassen et al., 2017; Klassen et al., 2018). In an 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 and their incremental validity over-and-above cognitive ability and personality tests (see 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 biases, 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. The present work therefore investigated key issues surrounding adverse impact in terms of gender, ethnicity, and socio-economic status (SES). 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 analyzed 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 (additive effects), 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 (multiplicative effects).

Using SJTs to assess prospective teachers' non-cognitive attributes

SJTs have been introduced to the teacher education domain as a way to improve the measurement of non-cognitive teacher attributes, such as motivation and personality, at the point of selection into initial teacher training (see Klassen & Kim, 2017, Klassen et al., 2018; Klassen et al., 2014 for an overview). A valid assessment of prospective teachers' non-cognitive attributes is critical, as non-cognitive attributes have consistently been linked to teaching performance (e.g., Kim, Jörg, & Klassen, 2019; Klassen & Tze, 2014) and might be even more important in that regard than cognitive attributes (e.g., Aloe & Becker, 2009; Bardach & Klassen, 2019; 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), because they more indirectly and implicitly assess applicants' judgments of (in)appropriate responses (Johnson & Saboe, 2011; Motowidlo & Beier, 2010).

While the teacher selection SJTs cover several domains of non-cognitive attributes (see Klassen et al., 2014; 2017; 2019), 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 the important characteristics of effective teachers (Klassen et al., 2019). 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 (e.g., Baier, Decker, Voss, Kleickmann, Klusmann, & Kunter, 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 have pervasive consequences for students' performance and self-beliefs (e.g., Timmerman, Kuyper, & van der Werf, 2015; Rubie-Davies, 2010; 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.

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., 2019), 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 a 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 and colleagues (2008), relates to the 'personality load' of a SJT, i.e., 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 and colleagues (2019) reported gender differences in scores on text-based SJTs (females > males). In a further study comparing different SJT formats, gender biases 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, 2019). 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 tends to be lower than those reported for cognitive ability tests (e.g., Whetzel & McDaniel, 2009). A prominent approach to explaining ethnicity effects in SJT performance focuses on the strength of the correlation between the SJT and cognitive ability tests ('cognitive load' of SJTs, Whetzel et al., 2008; also see e.g., Bobko, Roth, & Buster, 2005). Scores on measures of general cognitive ability are typically in favour of White participants, and hence, a SJT with a higher cognitive load should increase ethnic subgroup differences (Lievens et al., 2008; Whetzel et al, 2008). In a similar vein, Roth, Bobko, and Buster (2013) divided SJTs on a construct-level into cognitively-related scales and interpersonal scales and found that SJTs in the interpersonal category disadvantaged Black participants (vs. White participants) to a lesser extent than the cognitive ones.

The use of video-based SJTs has been discussed as another way to decrease ethnic group test score gaps (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., 2019). Still, 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 (see e.g., the findings of Roth et al., 2013). Although all of the SJT items developed for teacher selection represent challenging social situations, it is plausible that items targeting the domains of emotion regulation and mindset have an elevated cognitive saturation in comparison to items assessing conscientiousness: The latter should, to a higher degree, reflect individual differences in the personality trait of consciousness and might therefore be less cognitively loaded. This, in turn, could eventually lead to stronger ethnicity effects for the domains of 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, 2014). 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, calling for empirical investigations taking 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 (additive effects) of subgroup variables such as gender, ethnicity, and SES, to also investigating interactions (multiplicative effects) among these predictors (see e.g., Griffin & Hu, 2016).

The present study

The purpose of this study was to examine in more depth whether gender, ethnicity, and SES affect 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 additive and multiplicative effects of gender, ethnicity, and SES on overall SJT scores (*research question 1*). It was hypothesized that ethnicity and gender 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; Bardach et al., 2019), and that SES will affect SJT scores, favouring applicants from a higher SES background (e.g., Lievens et al., 2016). Moreover, we expected that gender would influence SJT performance, with females outperforming males (e.g., Klassen et al., 2019), given that the majority of SJT items employed in this study were text-based (see section on Measures). In addition, we proposed that there would be a significant interaction between gender and SES, gender and ethnicity, and SES and ethnicity, in that lower SES and being male should compound the effects of ethnic minority group membership disadvantage and that lower SES should compound the effect of gender disadvantages.

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 (*research question 2*). While we did not specify differentiated hypotheses for SES, we assumed that gender effects should emerge for all three domains, but should be stronger for conscientiousness than for mindset and emotion regulation (Whetzel et al., 2008). By contrast, the advantage of ethnic majority groups should be more pronounced for mindset and emotion regulation, as these domains might be more cognitively-related (e.g., Roth et al., 2013). Given that this study is the first to explore the potential value of studying separate domains, we did not feel confident enough to formulate specific hypotheses regarding interactions and conducted exploratory analyses in that regard.

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 program in the UK 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 UK (e.g., acceptable A-level exam results in relevant subjects and an undergraduate degree [at level 2:1 or better] in a relevant teaching subject, see Klassen et al., 2019 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, range restriction concerns were minimized, representing a considerable strength to the current study.

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, 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).

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 included 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., 2019), 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 (see Bergman, Drasgow, Donovan, Henning, & Juraska, 2006), whereby concordance panels with subject matter experts (SMEs) in the field determined the initial scoring key; however, subsequent revisions were made based upon the level of expert agreement, item quality, and the scoring patterns of the top ten percent 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 x 3 maximum points) equating to a total available score of 132 (11 scenarios x 12 maximum points). The reliability coefficients (Cronbach's α) for the SJT was α = .59, (consistent with mean Cronbach's αs for the SJT format [Campion, Ployhart, & MacKenzie, 2014]).

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, i.e., White background, 1 = minority group, i.e., participants from all other backgrounds), (3) home SES background (0 = high(er) SES status, i.e., those participants who indicated that they had neither received FSM nor EMA, 1 = low(er) SES status, i.e., 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., 2019) 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). We estimated two moderated regression models, one for each research question (research question 1: adverse impact on overall SJT performance, research question 2: adverse impact on SJT domain scores) relying 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). In the first model, we estimated additive effects (main effects) and multiplicative effects (interactions) of gender, SES, and ethnicity on overall SJT performance and in the second model, we explored additive and multiplicative effects on SJT performance for the three domains separately. We relied on manifest mean SJT scores in both models and the measures of gender, ethnicity, and SES consisted of single indicators.

We report unstandardized and regression standardized coefficients for the additive and multiplicative effects. 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. This approach takes all available information from the observed data into account when estimating parameter estimates and standard errors.

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.

[Insert Table 1 here]

The results of the model for research question 1 (subgroup differences in overall SJT performance) revealed no significant effect of gender (standardized $\hat{\beta}$ = -0.042, p = .051) and no significant effect of SES (standardized $\hat{\beta}$ = -0.007, p = .413) on SJT scores. Ethnicity significantly predicted SJT performance, with members from ethnic majority groups performing better than members from minority groups (standardized $\hat{\beta}$ = -0.282, p < .001). Moreover, the interaction terms between gender and ethnicity proved significant (standardized $\hat{\beta}$ = -0.106, p < .001). While SJT scores were higher for members from majority groups than for members from minority groups, this effect was more pronounced in male ethnic minority group members (see Figure 1). The results also showed a significant positive interaction between ethnicity and SES in predicting SJT scores (standardized $\hat{\beta}$ = 0.093, p = .007). This means that ethnic minority group members stemming from a low(er) SES background performed better than those from a high(er) SES background (but still scored lower than ethnic majority group members irrespective of their SES background, see Figure 1). The interaction between gender and SES was not statistically significant (standardized $\hat{\beta}$ = 0.013, p = .345).

The analyses that were conducted for the separate domains indicated that ethnicity significantly predicted SJT scores in the domain of conscientiousness (standardized $\hat{\beta}$ = -0.129, p < .001, with ethnic majority group members > ethnic minority group members). There were no significant effects of gender and SES on conscientiousness scores (standardized $\hat{\beta}$ = -0.040, p = .070, and standardized $\hat{\beta}$ = 0.016, p = .305, respectively). The interactions between gender and SES, as well as ethnicity and SES, did not reach statistical significance (standardized $\hat{\beta}$ = -0.026, p = .435, standardized $\hat{\beta}$ = 0.044, p = .223, respectively). However, the results yielded a significant interaction between ethnicity and gender (standardized $\hat{\beta}$ = -0.071, p = .028). Mirroring the results for overall SJT scores, males from ethnic minority groups were most disadvantaged.

With regard to the domain of mindset, the effects of gender and SES on mindset SJT scores were not statistically significant (standardized $\hat{\beta}$ = 0.016, p = .270 for gender and standardized $\hat{\beta}$ = -0.016, p = .320 for SES), but ethnicity significantly predicted SJT scores (favouring members from ethnic majority groups, standardized $\hat{\beta}$ = -0.210, p < .001). The interaction between gender and SES was not statistically significant (standardized $\hat{\beta}$ = 0.009, p = .787), but we found a statistically significant interaction between gender and ethnicity (standardized $\hat{\beta}$ = -0.111, p = .001). Members from ethnic minority groups did not perform as well as members from ethnic majority groups and the size of the effect was larger for males than females (see Figure 1). Finally, we found a significant interaction between ethnicity and SES (standardized $\hat{\beta}$ = 0.082, p = .027). While members from ethnic minority groups did less well on the mindset SJTs than members from ethnic minority groups, this disadvantage affected those from a high(er) SES background more than those from a low(er) SES background (see Figure 1).

For emotion regulation, a significant negative effect of gender was found, with females outperforming males (standardized $\hat{\beta}$ = -0.062, p = .008) and a significant negative effect of ethnicity, with members from ethnic majority groups outperforming ethnic minority group members (standardized $\hat{\beta}$ = -0.257, p < .001). SES did not significantly predict emotion regulation SJT scores (standardized $\hat{\beta}$ = -0.014, p = .327). None of the interaction terms were statistically significant (standardized $\hat{\beta}$ = 0.042, p = .202 for the interaction between gender and SES, standardized $\hat{\beta}$ = -0.044, p = .198 for the interaction between gender and ethnicity, and standardized $\hat{\beta}$ = 0.071, p = .061 for the interaction between ethnicity and SES). Table 2 shows all standardized and unstandardized regression coefficients for the main effects and interactions, and Figure 1 and Figure 2 illustrate the interactions for SJT composite and domain scores, respectively.

[Insert Table 2 here]

[Insert Figure 1 here]

[Insert Figure 2 here]

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, 2014), 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 additive and multiplicative effects and relying on overall SJT scores as well as domain-specific scores. With regard to additive effects on SJT overall scores, we found that males did not experience a significant adverse impact. This result is in contrast to meta-analytic findings (Whetzel et al., 2008) and the study by Klassen and colleagues (2019) with a text-based teacher selection SJT. It partially aligns with the findings of Bardach and colleagues (2019) who showed that no gender differences emerged in conditions with video-based SJT items as compared to a text-based SJT condition. Whether the absence of gender effects in our study might thus be due to the fact that several 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 spill-over 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 at least a small number of video-SJTs included.

Furthermore, our results indicated the known effects of ethnicity on SJT scores, whereby members from ethnic majority groups outperform members from ethnic minority groups (e.g., Whetzel et al., 2008). This finding raises 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 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 (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, 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).

Analyzing the interplay between the three individual difference variables in forecasting SJT scores revealed that the interaction between gender and SES was not significant. However, the main effects for ethnicity have to be reconsidered in light of the fact that ethnicity significantly interacted with both gender and SES, and hence, the fact that ethnicity effects vary depending on gender and SES characteristics. Confirming our hypothesis concerning the interaction between ethnicity and gender, the disadvantage of ethnic minority group membership was exacerbated for applicants identifying as male. Surprisingly, the results for the second significant interaction between ethnicity and SES did not demonstrate, as hypothesized, that being from a low(er) SES background may increase performance discrepancies. Instead, the findings indicated that ethnic minority group members from a high(er) SES background obtained lower SJT scores. Of course, this finding might be context-related and thus, specific to this sample and this study. Nonetheless, we want to stress a further possible explanation. It could be that ethnic minority group members from a low(er) SES background who strive to become teachers might be particularly motivated to do well on the test and put in a lot of effort, maybe also because they are aware of the stereotypes related to the 'double disadvantage' they face.

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 showed that (small) gender effects were restricted to emotion regulation, thus contradicting our hypothesis that conscientiousness would be the domain most strongly impacted by gender effects (e.g., Whetzel et al., 2008). In accordance with the findings for SJT overall scores, low(er) SES background was not related to performance decrements in any of the domains. However, the pattern of findings concerning the impact of ethnicity confirmed our predictions that stronger ethnicity effects may occur for the potentially more cognitively-focused domains of mindset and emotion regulation than for conscientiousness (Roth et al., 2013). In addition, a significant interaction between gender and ethnicity emerged for conscientiousness and mindset, which mirrored the interaction found for SJT total scores, with males from ethnic minority groups receiving the lowest scores. The ethnicity-SES interaction was solely significant for mindset and none of the interactions for emotion regulation proved significant, pointing towards the potential of studying separate domains to more thoroughly understand adverse impact. For mindset, the significant interaction between ethnicity and SES revealed that members from ethnic minority groups could partially compensate for their ethnicity-related disadvantage if they came from low(er) SES backgrounds. This is a surprising finding and, as we have outlined above when interpreting this interaction for SJT overall scores, one possibly confounding factor could be the higher motivation or better preparation of ethnic minority members from a low(er) SES background in our sample that allowed them to catch up a bit.

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 program 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. For example, only 13 applicants identified as non-binary, making valid comparisons between this category and the currently included categories male vs. female difficult.

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., 2019), they also extend thinking about adverse impact 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.

References

Aloe, A. M., & Becker, B. J. (2009). Teacher verbal ability and school outcomes: where is the evidence? *Educational Researcher*, *38*, 612-624.

Albert Shanker Institute. (2015). *The state of teacher diversity in American education.* Washington, DC: Author. Retrieved from http://www.shankerinstitute.org/resource/teacherdiversity.

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*. doi:10.1111/bjep.12256

Bardach, L., Rushby, J.V., Kim, L.E., & Klassen, R.M. (2019). Using video-based situational judgment tests for teacher selection: A quasi-experiment exploring the relations between test format, subgroup differences, and applicant reactions. *PsyArXiv.* June, 20.

Bardach, L., & Klassen, R.M. (2019). Smart teachers, successful students? A systematic review of the literature on teachers' cognitive abilities and teacher effectiveness. *PsyArXiv.* March, 28.

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.

Bobko, P., Roth, P. L., & Buster, M. A. (2005). Work sample selection tests and expected reduction in adverse impact: A cautionary note. *International Journal of Selection and Assessment, 13*, 1-10.

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.

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.

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.

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.

Crosnoe, R., & Muller, C. (2014). Family socioeconomic status, peers, and the path to college. *Social Problems*, *61*(4), 602-624.

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.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford.

Festin, 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.

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.

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.

Griffin, B., & Hu, W. (2015). The interaction of socio-economic status and gender in widening participation in medicine. *Medical Education*, *49*(1), 103-113.

Johnson, R., & Saboe, K. (2010). Measuring Implicit Traits in Organizational Research: Development of an Indirect Measure of Employee Implicit Self-Concept. *Organizational Research Methods*, *14*(3), 530-547.

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 (3),* 301–323.

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.

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.

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 in Education*, 96-114.

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.

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*. Oxford: Wiley.

Klassen, R., & Tze, V. M. C. (2014). Teachers' self-efficacy, personality, and teaching effectiveness: A meta-analysis. *Educational Research Review*, *12*, 59-76.

Lievens, F. (2013). Adjusting medical school admission: assessing interpersonal skills using situational judgement tests. *Medical Education*, *47*(2), 182-189.

Lievens, F. (2017). Construct-driven SJTs: Toward an agenda for future research. *International Journal of Testing, 17*(3), 269-276.

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*(6), 624-636.

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, *37*(4), 426-441.

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.

Muthén, L.K. and Muthén, B.O. (1998-2010). *Mplus User's Guide. Sixth Edition*. Los Angeles, CA: Muthén & Muthén.

Muthén, B. O., Muthén, L. K. & Asparouhov, T. (2016). *Regression and Mediation Analysis Using Mplus*. Los Angeles, CA: Muthén & Muthén.

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.

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.

OECD (2016). *Teachers, the Learning Environment and the Organisation of Schools.* Paris, France: OECD Publishing.

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.

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*(5), 532-550.

Patterson, F., Ashworth, V., & Good, D. (2013). Situational judgement tests: A guide for applicants to the UK Foundation Programme. *Medical Schools Council*, 1-29.

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*(3), 394-409.

Timmermans, A. C., Kuyper, H., & van der Werf, G. (2015). Accurate, inaccurate, or biased teacher expectations: Do D utch teachers differ in their expectations at the end of primary education? *British Journal of Educational Psychology, 85*(4), 459-478.

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*(6), 740-746.

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review*, *19*(3), 188-202.

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance*, *21*(3), 291-309.

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*(5), 648-668.

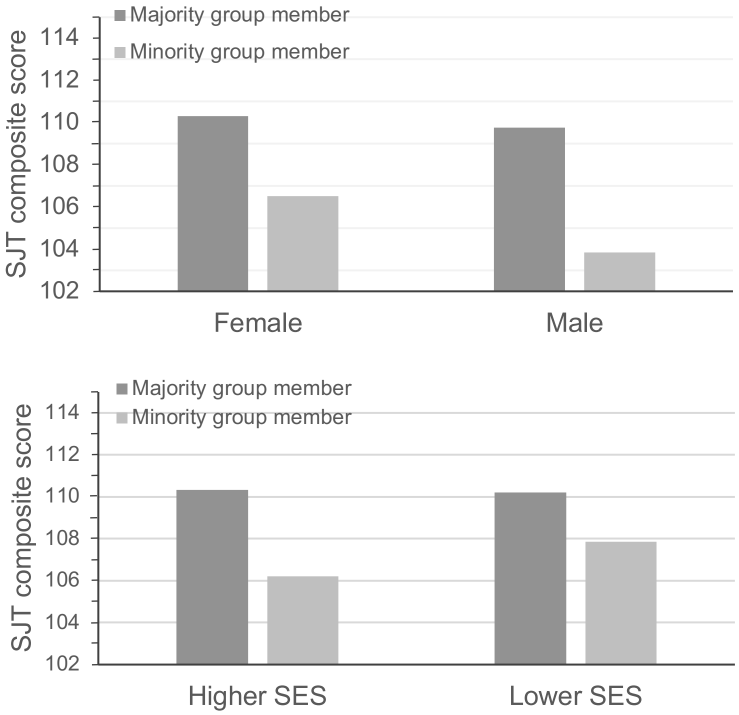

Figure 1. The interaction between ethnicity and gender, and ethnicity and socio-economic status (SES) on Situational Judgment Test (SJT) performance

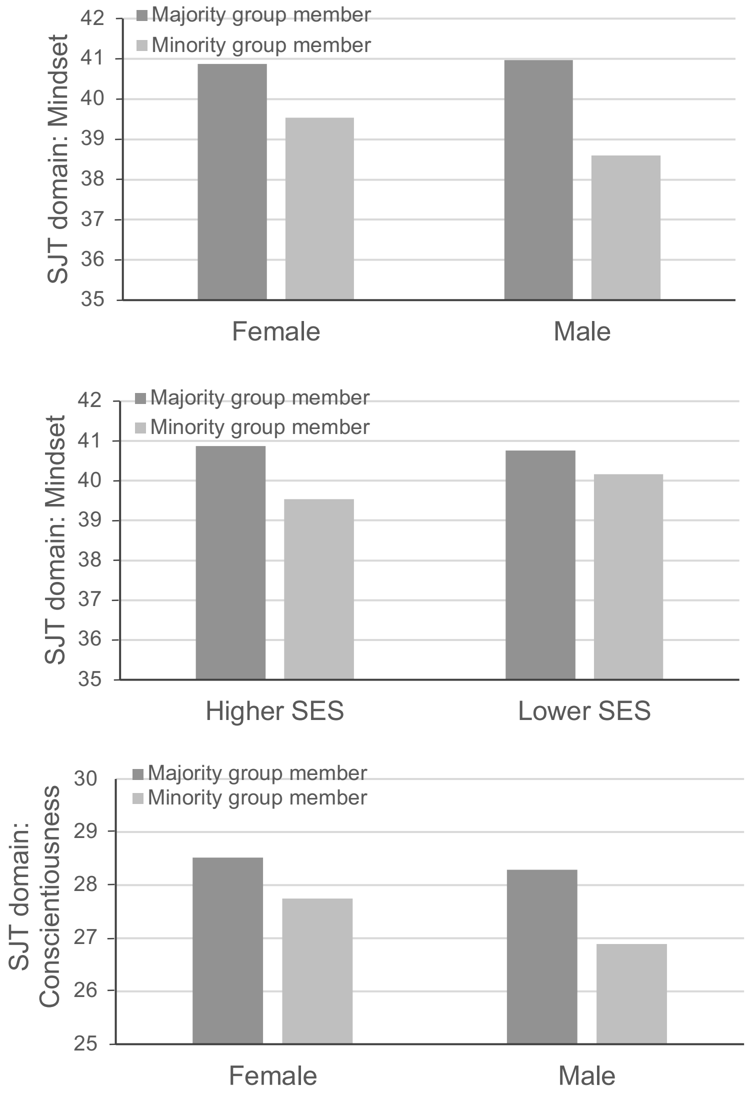

Figure 2. The interaction between ethnicity and gender, and ethnicity and socio-economic status (SES) on Situational Judgment Test (SJT) domain performance

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**	**-.09**	-.03	**-.06**			
6. Ethnicity	-.29	**-.14**	**-.22**	**-24.**	-.00		
7. SES	.05	-.00	-.02	-.01	-.01	**.20**	
M	108.89	28.71	40.43	40.28			
SD	6.29	2.77	3.00	3.14			

Note: SJT = Situational Judgment 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 = lowe(er) SES background; Statistically significant correlation coefficient at α = .05 are boldface.

Table 2

Unstandardized and Standardized Estimates of all Effects

Effects	Unstandardized estimates (*S.E.*)	Standardized estimates (*S.E.*)
SJT composite score		
Gender → SJT scores	-0.088 (0.054)	-0.042 (0.026)
Ethnicity → SJT scores	**-0.606 (0.072)**	**-0.282 (0.032)**
SES → SJT scores	-0.016 (0.072)	-0.007 (0.033)
Gender x Ethnicity → SJT scores	**-0.335 (0.110)**	**-0.106 (0.035)**
Gender x SES → SJT scores	0.045 (0.112)	0.013 (0.033)
Ethnicity x SES → SJT scores	**0.280 (0.114)**	**0.093 (0.013)**
Domains: Conscientiousness		
Gender → SJT scores	-0.084 (0.057)	-0.040 (0.027)
Ethnicity → SJT scores	**-0.278 (0.067)**	**-0.129 (0.031)**
SES → SJT scores	0.036 (0.070)	0.016 (0.032)
Gender x Ethnicity → SJT scores	**-0.224 (0.102)**	**-0.071 (0.032)**
Gender x SES → SJT scores	-0.086 (0.111)	-0.026 (0.033)
Ethnicity x SES → SJT scores	0.131 (0.108)	0.044 (0.036)
Domains: Mindset		
Gender → SJT scores	0.034 (0.055)	0.016 (0.027)
Ethnicity → SJT scores	**-0.449 (0.073)**	**-0.210 (0.033)**
SES → SJT scores	-0.036 (0.077)	-0.016 (0.035)
Gender x Ethnicity → SJT scores	**-0.350 (0.110)**	**-0.111 (0.035)**
Gender x SES → SJT scores	0.031 (0.114)	0.009 (0.034)
Ethnicity x SES → SJT scores	**0.247 (0.113)**	**0.082 (0.037)**
Domains: Emotion Regulation		
Gender → SJT scores	**-0.130 (0.054)**	**-0.062 (0.026)**
Ethnicity → SJT scores	**-0.552 (0.076)**	**-0.257 (0.034)**
SES → SJT scores	-0.031 (0.069)	-0.014 (0.031)
Gender x Ethnicity → SJT scores	-0.140 (0.109)	-0.044 (0.034)
Gender x SES → SJT scores	0.143 (0.112)	0.042 (0.033)
Ethnicity x SES → SJT scores	0.213 (0.114)	0.071 (0.038)

Note. Moderated regression results. *S.E.* =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 = lowe(er) SES background; Statistically significant results at α = .05 are boldface.