Received: 12 June 2019 | Revised: 13 September 2019 | Accepted: 17 September 2019

DOI: 10.1111/ijsa.12269

FEATURE ARTICLE

International Journal of Selection and Assessment WILEY

Subgroup differences in situational judgment test scores: Evidence from large applicant samples

Christoph N. Herde[1] | Filip Lievens[2] | Duncan J. R. Jackson[3,4] | Ali Shalfrooshan[5] | Philip L. Roth[6]

[1]Department of Personnel Management, Work and Organizational Psychology, Ghent University, Ghent, Belgium

[2]Lee Kong Chian School of Business, Singapore Management University, Singapore, Singapore

[3]King's Business School, King's College London, London, UK

[4]Department of Industrial Psychology, University of the Western Cape, Bellville, South Africa

[5]PSI Talent Measurement, Guildford, UK

[6]Department of Management, College of Business, Clemson University, Clemson, South Carolina

Correspondence
Christoph N. Herde, Department of Personnel Management, Work & Organizational Psychology, Ghent University, Henri Dunantlaan 2, Ghent 9000, Belgium.
Email: christoph.herde@ugent.be

Abstract

To promote diversity in organizations it is important to have accurate knowledge about subgroup differences associated with selection procedures. However, current estimates of subgroup differences in situational judgment tests (SJTs) are overwhelmingly based on range-restricted incumbent samples that are downwardly biased. This study provides much-needed applicant level estimates of SJT subgroup differences (N = 37,530). As a key finding, Black-White differences (d = 0.66) were higher than in incumbent samples (d = 0.38). Overall, sex differences were small. Females scored higher for management jobs (d = −0.13) and males scored higher for administrative jobs (d = 0.15). By analyzing applicant samples that do not suffer from range restriction, this study adds knowledge about subgroup differences in SJTs.

1 | INTRODUCTION

Twenty-first-century organizations are involved in a global search for talent and typically select from applicant pools that differ in terms of their ethnic makeup (Cascio & Aguinis, 2008). Therefore, it is important for organizations to have accurate knowledge about the extent to which personnel selection procedures exhibit subgroup differences (Ployhart & Holtz, 2008). That is, if existing selection procedures produce substantial mean differences between applicants of varying subgroups based on demographic variables (such as ethnicity or sex), it may be more difficult for specific (minority) subgroups to pass selection stages and to start working in organizations, thereby impeding the development of diversity in organizations (Arthur, Edwards, & Barrett, 2002; Ployhart & Holtz, 2008; Sackett, Schmitt, Ellingson, & Kabin, 2001).

Although there exists a substantial base of research about the degree to which scores on selection procedures are generally associated with subgroup differences, a key limitation is that most of the effect sizes are based on *incumbent* samples (Roth, Bobko, Switzer, & Dean, 2001). Given that incumbents have typically gone through an extensive selection process prior to entering the organization and are thus range-restricted, effect sizes on the basis of incumbent samples tend to be downwardly biased (Bobko & Roth, 2013). Therefore, this study provides much-needed large-scale, *applicant*-level estimates of subgroup differences related to one specific selection procedure: situational judgment tests (SJTs).

2 | STUDY BACKGROUND

2.1 | Subgroup differences in selection

Over the past years, this need for accurate information about subgroup differences in personnel selection procedures has generated

Parts of this paper were presented at the 32nd Annual Conference of the Society for Industrial and Organizational Psychology (SIOP), Orlando, Florida.

Ali Shalfrooshan is the Head of International Assessment R&D at a company that sells selection and development procedures, including SJTs.

a substantial amount of research, resulting in several narrative summaries and meta-analyses with special attention given to subgroup differences across ethnicities (Bobko & Roth, 2013; Dean, Roth, & Bobko, 2008; Hough, Oswald, & Ployhart, 2001; Huffcutt & Roth, 1998; Lynn & Irwing, 2004; Ployhart & Holtz, 2008; Roth, Bevier, Bobko, Switzer, & Tyler, 2001; Roth, Bobko, McFarland, & Buster, 2008; Roth, Van Iddekinge, et al., 2017; Whetzel, McDaniel, & Nguyen, 2008).

In an extensive review of past research on ethnic subgroup differences between Black and White test-takers, Bobko and Roth (2013) summarized the main conclusions about the most prominent selection procedures. In comparison to other selection procedures, general cognitive ability tests produce Black–White differences of about *d* = 1.00 in many situations in society, and mean *d*'s of 0.72 and 0.86 for medium and low complexity jobs, respectively. These values of *d* = 0.72, 0.86, and 1.00 are then typically used as a benchmark against which Black–White subgroup differences are compared for other selection procedures. For example, Bobko and Roth (2013) reported that work samples and structured interviews were generally associated with relatively lower Black–White differences, with values such as of *d* = 0.38 and *d* = 0.25.

However, in their review, Bobko and Roth (2013) pointed to a key limitation that challenges previously reported subgroup differences for various selection procedures. That is, they argued that past empirical evidence provides downwardly biased information for decision-makers in organizations because the majority of primary studies of alternative predictors (e.g., work sample tests, SJTs) involved incumbent samples which suffer from range restriction (see also Roth, Le, Oh, Van Iddekinge, & Robbins, 2017).[1] Incumbents have been selected for their job either by (a) the respective personnel selection procedure that is under scrutiny (i.e., direct range restriction) or (b) other selection procedures that are correlated with the focal selection procedure to be analyzed (i.e., indirect range restriction). Therefore, lower scoring job applicants are less likely to be represented in incumbent samples and are more likely to be represented in applicant samples. Hence, (a) direct or (b) indirect range restriction may lead to estimates of subgroup differences that are substantially underestimated when incumbent samples are analyzed (Bobko & Roth, 2013).

Bobko and Roth (2013) provided empirical support for their arguments by comparing published effect sizes for Black–White mean differences based on range-restricted, incumbent samples with limited available applicant data. In personnel selection procedures such as work samples and interviews, Black–White mean differences favoring Whites in applicant samples were higher than in previous range-restricted incumbent samples. Specifically, Bobko and Roth (2013) concluded that estimates of Black–White differences in incumbent samples equaled 0.38 for work samples and 0.25 for structured interviews. For applicant samples, however, Bobko and Roth reported higher effect sizes of 0.73 for work samples and between 0.31 and 0.46 for structured interviews. For other selection methods, there is less clarity in this respect and thus the actual size of Black–White differences remains unclear (Bobko & Roth, 2013). This is particularly notable in the case of SJTs which we describe below.

2.2 | Subgroup differences and SJTs

In their traditional format, SJTs present applicants with written descriptions of job-related situations and different predetermined response options. Applicants then need to choose one of the response options, rank them, or rate each of them in terms of their presumed effectiveness or typicality for the applicants' behavior (Motowidlo, Dunnette, & Carter, 1990). Depending on the domain of interest (interpersonal, leadership, etc.), SJTs assess applicants' procedural knowledge about appropriate (interpersonal, leadership, etc.) behavior on the job (Lievens, 2017; Lievens & Motowidlo, 2016; Motowidlo & Beier, 2010; Motowidlo, Hooper, & Jackson, 2006). SJTs are frequently used in selection due to the evidence for adequate criterion-related validity (Christian, Edwards, & Bradley, 2010; McDaniel, Hartman, Whetzel, & Grubb, 2007) and favorable applicant perceptions (Kanning, Grewe, Hollenberg, & Hadouch, 2006).

Current knowledge about ethnic subgroup differences in SJTs mainly builds upon the meta-analysis of Whetzel et al. (2008). This study showed that SJTs consistently favor Whites in comparison to other ethnicities. Standardized mean subgroup differences equaled *d* = 0.38 for Black–White, *d* = 0.24 for Hispanic–White, and *d* = 0.29 for Asian–White comparisons. Whetzel et al. (2008) further identified the cognitive saturation (i.e., the extent to which cognitive ability influences a measure; see Lubinski & Dawis, 1992) of SJTs as an important moderator of these ethnic subgroup differences. That is, the cognitive saturation of SJTs was positively related to the extent of ethnic subgroup differences. The relation between the cognitive ability saturation and ethnic subgroup differences equaled *r* = .77 for Black–White differences, *r* = .49 for Hispanic–White differences, and *r* = .40 for Asian–White differences.

Whetzel et al. (2008) also reported sex group differences for SJTs. Their overall results indicated that females, on average, scored higher than males (mean *d* = −0.11). Thus, sex differences were quite small for SJTs and indicate that adverse impact potential against females is not generally likely. Whetzel et al. were also explicit about the fact that virtually all of their data were obtained from incumbent samples for both ethnic and sex group differences.

The distinction between incumbent versus applicant samples has important implications. According to Bobko and Roth (2013, p. 104), "More accurate *d* values for SJTs are unclear, though likely larger than currently expected because of dependence on incumbent samples." In that vein, Roth, Bobko, and Buster (2013) is the only study that examined SJT subgroup differences in applicant samples. They found that effect sizes for Black–White differences varied indeed between applicant samples. Roth et al. showed that Black–White mean differences in SJT scores reached values of *d* = 0.67 and *d* = 0.61, for jobs in which Blacks were the minority. Yet, these results were based on small minority sample sizes of *N* of approximately 50 and limited total sample sizes (*N* of approximately 1,200, see Roth et al., 2013). In any case, these results underscore the importance of both increasing the sample sizes of minorities and of including a wider range of jobs in applicant samples. Roth et al. did also not report data for sex group differences. Thus, applicant level differences on this variable are virtually unexplored in the current literature.

Taken together, the research base on Black–White subgroup differences in SJT scores at the applicant level is quite limited. Therefore, decision-makers in organizations still likely rely on the meta-analytic effect sizes based on incumbent samples for benchmarking subgroup differences found for SJT scores. To improve our knowledge of ethnic subgroup differences in SJTs at the applicant level (see Bobko & Roth, 2013; Whetzel et al., 2008), more primary studies based on applicant samples are needed. Importantly, those studies should also further analyze the nature of subgroup differences in applicant scores in different job domains. Finally, past applicant samples mainly incorporated Whites and Blacks and, hence, they did not contribute to our knowledge related to other ethnic subgroups such as Asians, a group that has received notably less attention than other minority groups (see Bobko & Roth, 2013; Lievens & De Soete, 2012; Roth, Van Iddekinge, et al., 2017). Similarly, we could find no applicant studies of sex differences on SJT scores in the applied psychology literature. As such, researchers and decision-makers have little empirical guidance in this area too.

3 | PRESENT STUDY

This study aims to extend knowledge relating to ethnic and sex subgroup differences in SJT scores at the *applicant* level. Specifically, we present large-scale applicant samples to compare overall SJT scores between Whites, Blacks, and Asians, as well as males and females. Our focus on applicant samples avoids range restriction issues that are typical in incumbent samples. In line with prior studies on subgroup differences related to SJT scores (e.g., Whetzel et al., 2008), we also report results for comparisons between females and males. Moreover, we analyze data based on SJTs used in selection processes for different target job families.

4 | METHOD

4.1 | Participants

We report data provided by a business consultancy that developed the SJTs under scrutiny and offered them to organizations to assist in guiding selection decisions. We received data on four independent applicant samples. Each of these four samples consists of applicants for a specific job family in organizations located in the United Kingdom.[2] Although the selection process of each sample shared the same job family and was very similar (see Procedure), the samples pooled applicants from a variety of organizations and industry sectors. Among others, industry sectors included the public, financial services, retail, manufacturing, and health-care sectors.

The first sample (administrative sample) comprised 5,047 applicants (mean age = 31.23 years, *SD* = 10.74 years) for administrative, clerical, and secretarial staff. The second sample (customer service sample) consisted of 5,592 applicants (mean age = 29.27 years, *SD* = 11.15 years) for customer service staff. The third sample (graduate sample) included 18,579 applicants (mean age = 26.68 years, *SD* = 7.89 years) for graduate positions. These graduate positions were typically entry level jobs for applicants who completed at least a 3-year university degree. These jobs did usually not require specific expertise or work experience. The fourth sample (management sample) was composed of 8,312 applicants (mean age = 39.01 years, *SD* = 9.81 years) for first line or middle management roles. Information about applicants' ethnicity and sex by sample are provided in Tables 1 and 2.

TABLE 1 Ethnic group differences for job applicants by situational judgment test

Sample	White (W)	Black (B)	$d_{\text{W-B}}$	95% CI	Asian (A)	$d_{\text{W-A}}$	95% CI
Management	95.80 (9.67) *n* = 6,157	89.80 (11.37) *n* = 478	0.61	[0.52; 0.71]	89.60 (12.80) *n* = 700	0.62	[0.54; 0.70]
Graduate	87.58 (10.72) *n* = 8,377	78.91 (11.93) *n* = 2,343	0.79	[0.74; 0.84]	81.35 (11.62) *n* = 4,299	0.56	[0.53; 0.60]
Customer service	80.90 (11.03) *n* = 5,082	74.57 (11.84) *n* = 136	0.57	[0.40; 0.74]	75.59 (12.37) *n* = 203	0.48	[0.34; 0.62]
Administrative	59.00 (7.53) *n* = 3,748	55.25 (7.48) *n* = 383	0.50	[0.39; 0.60]	55.45 (8.15) *n* = 565	0.47	[0.38; 0.56]
Overall	*n* = 23,364	*n* = 3,340	0.66		*n* = 5,767	0.55	

Note: Mean overall scores appear in columns relating to ethnic origin. Standard deviations appear in parentheses. Standard differences (*d*) are presented using the White category as the referent (e.g., $d_{\text{W-B}}$ = standardized difference between White and Black groups). Overall standard difference was computed as weighted average across the four samples.

TABLE 2 Sex Differences for job applicants by situational judgment test

Sample	Male (M)	Female (F)	$d_{M\text{-}F}$	95% CI
Management	94.20 (10.55) *n* = 4,405	95.53 (10.23) *n* = 3,369	−0.13	[−0.17;−0.08]
Graduate	84.47 (11.76) *n* = 9,108	83.89 (11.94) *n* = 6,894	0.05	[0.02; 0.08]
Customer service	80.28 (11.35) *n* = 2,827	80.78 (11.01) *n* = 2,685	−0.04	[−0.10; 0.01]
Administrative	58.96 (7.71) *n* = 1,927	57.81 (7.74) *n* = 2,994	0.15	[0.09; 0.21]
Overall	*n* = 18,267	*n* = 15,942	0.01	

Note: Mean overall scores appear in columns relating to sex grouping. Standard deviations appear in parentheses. Standardized differences ($d_{M\text{-}F}$) are presented for each test. Overall standard difference was computed as weighted average across the four samples.

4.2 | Procedure

The selection process for all of the four samples was similar in terms of structure and included three selection stages. The first stage included a basic background check to determine whether applicants fulfilled the minimum qualification as stated in the job advertisement (e.g., minimum level of education). This check did not resemble a detailed background check that includes in-depth analyses of applicants' resumes, criminal records, or credit history.[3] At the second stage, applicants took part in an online screening process in an unproctored setting. Importantly, across all four samples, this online screening process included an SJT. Organizations that used the SJTs to screen applicants were advised to remove candidates who scored below the 30th percentile in the SJT. Depending on the knowledge, skills, abilities, and other characteristics needed for the job family, applicants also completed additional tests. Data for these additional tests were not available for our analyses. At the third stage, applicants participated in a job interview and/or in an assessment center.

4.3 | SJT

The content of the SJTs differed across the four independent samples. The written SJTs were developed in accordance with professional guidelines and approaches outlined in other studies (e.g., Weekley, Ployhart, & Holtz, 2006). First, a thorough analyses of the job families was conducted. Next, a team of three to five psychologists wrote scenarios on the basis of critical incidents gathered from incumbent representatives of each job family. Each scenario was followed by four item options. The development of these response options was also based on information gathered via critical incidents and the job analysis and was conducted by the same team of psychologists who wrote the scenarios. Afterward, distinct teams of seven to nine psychologists conducted another review of scenarios and response options.

For each scenario, knowledge-based instructions ("What should you do?") required applicants to independently rate the effectiveness of the four possible response options (items) on a 5-point Likert scale (1 = *counterproductive*, 5 = *very effective*). Given that the SJTs were used in a high-stakes selection setting, a knowledge-based response instruction format was chosen because this format is less prone to faking (Lievens, Sackett, & Buyse, 2009; Nguyen, Biderman, & McDaniel, 2005). However, note that knowledge-based response instructions showed slightly higher ethnic and sex group differences in past studies (Whetzel et al., 2008).

Across the four different SJTs, 80–101 subject matter experts (e.g., line managers) completed the items to determine the scoring key. In line with the *consensus weighting method* (see Chan & Schmitt, 1997), the relative frequency of subject matter expert endorsement of the specific points on the Likert scale determined the scoring key of 0, 1, or 2 per item. Response options endorsed by more than 50% of subject matter experts indicated a score of 2, response options endorsed by 25%–50% of subject matter experts indicated a score of 1, and response options endorsed by less than 25% of subject matter experts indicated a score of 0 (see Chan & Schmitt, 1997). All four SJTs were linear (i.e., nonbranched). Thus, candidates' prior responses to scenarios or response options did not change the content, order or number of scenarios or response options subsequently presented. Table A1 in the appendix provides an overview of number of scenarios and example scenarios for the four SJTs.

Test manuals for the SJTs report evidence in support of their validity. For example, the performance in the SJT used in the customer service sample relates positively to line manager ratings of performance ($r = .28$, $p < .01$). Further, the SJT used in the graduate sample shows an expected relation to overall performance in assessment centers ($r = .54$, $p < .01$) (e.g., Lievens & Patterson, 2011). Finally, performance in the SJT used in the management sample discriminates between incumbents of different management levels with higher scores in the SJT being related to higher management levels.

4.4 | Analyses

We examined standardized ethnic and sex subgroup differences in SJTs by estimating Cohen's *d* for the comparisons between Whites, Blacks, and Asians as well as females and males. While Whites served as reference group for ethnic comparisons, males represented the reference group for sex comparisons. For all of the four samples, standardized subgroup differences were estimated based on overall SJT scores. To compare standardized subgroup differences across samples as well as across Black–White and White–Asian differences, we calculated 95% confidence intervals. Finally, we computed overall weighted standard differences as weighted average across standard differences of the four samples.

5 | RESULTS

Overall SJT mean scores, standard deviations and coefficients alpha are shown in Table 3. Coefficients alpha for overall SJT scores ranged between α = .54 (administrative sample) and α = .77 (customer service sample).

5.1 | Ethnic subgroup differences

Overall weighted average standardized subgroup differences across all samples were estimated at *d* = 0.66 for Black–White comparisons. Cohen's *d* values indicated moderate, or slightly larger, subgroup differences favoring Whites for all four samples (see Table 1). Cohen's *d* estimates varied between *d* = 0.50 (administrative sample; 95% CI = [0.39; 0.60]) and *d* = 0.79 (graduate sample; 95% CI = [0.74; 0.84]). Inspection of the confidence intervals around the *d*s suggested significantly higher Black–White differences in the graduate sample as compared to the management and administrative samples. Overlapping confidence intervals suggested comparable Black–White differences across samples for all other comparisons. The lower bound of each sample confidence interval was above the meta-analytic Black–White mean difference of 0.38 noted above for incumbent-based samples (Whetzel et al., 2008).

Regarding comparisons between Whites and Asians, overall weighted average standardized subgroup differences across all samples were estimated at *d* = 0.55. Cohen's *d* values showed small to moderate subgroup differences between *d* = 0.47 (administrative sample; 95% CI = [0.38; 0.56]) and *d* = 0.62 (management sample; 95% CI = [0.54; 0.70]) favoring Whites (see Table 1). Confidence intervals for all effect sizes overlapped, indicating somewhat comparable White–Asian differences across job families. Notably, the lower bounds of all the confidence intervals were larger than the point estimate of 0.29 from Whetzel et al. (2008).

TABLE 3 Means, standard deviations and coefficients alpha by situational judgment test

Sample	*M*	*SD*	α
Administrative	58.23	7.74	.54
Customer service	80.54	11.18	.77
Graduate	84.24	11.69	.68
Management	94.76	10.41	.67

Inspection of confidence intervals for ethnic subgroup differences within each sample showed similar effect sizes for Black–White and White–Asian differences. As the only exception, Black–White differences were notably higher than White–Asian differences in the graduate sample (*d* = 0.79; 95% CI = [0.74; 0.84] vs. *d* = 0.56; 95% CI = [0.53; 0.60]).

5.2 | Sex subgroup differences

For comparisons between males and females, overall weighted average standardized differences across all four samples were estimated at *d* = 0.01. Cohen's *d* values indicated negligible-to-small subgroup differences (see Table 2). Effect sizes ranged between *d* = −0.04 (customer service sample; 95% CI = [−0.10; 0.01]) and *d* = 0.15 (administrative sample; 95% CI = [0.09; 0.21]). Males outperformed females in the administrative and graduate sample, whereas females scored higher in the management sample. As the associated confidence interval included zero, there were no notable sex differences related to the customer service SJT scores.[4]

6 | DISCUSSION

This study provides much-needed large-scale estimates of SJT subgroup differences at the applicant level (total *N* = 37,530). The subgroup differences dealt with Whites, Blacks, and Asians as well as with sex differences. These analyses were based on independent applicant samples for four different job families. Across all samples, Whites outperformed Blacks and Asians by at least around half a standard deviation. More precisely, our overall weighted standardized Black–White subgroup differences in SJT scores (*d* = 0.66) are higher as compared to the average meta-analytic estimate based on prior studies that included incumbent samples (*d* = 0.38; Whetzel et al., 2008).[5] The subgroup differences associated with SJT scores in applicant samples follow the same trend that Bobko and Roth (2013) outlined for other personnel selection procedures such as work samples or structured interviews. That is, applicant samples generate substantially higher estimates of ethnic subgroup differences as compared to samples including incumbents, because the latter suffer from (direct or indirect) range restriction that leads to downwardly biased effect sizes.

Apart from this key result, some other findings are also noteworthy. First, Black–White differences were significantly higher in the graduate sample as compared to the management and administrative samples. On the basis of the available data sets, it is difficult to posit conclusive explanations for this result. The SJTs were not only taken by different applicants but also differed in terms of job family and content. Accordingly, the items might vary in their construct saturation. We were unable to investigate the construct saturation of the SJT items because we did not have access to item content or data. Bobko and Roth (2013) showed that many selection methods exhibit Black–White differences because of specific constructs being targeted. In line with Bobko and Roth (2013), we therefore welcome future research to compare subgroup differences in SJTs by respective target construct (e.g., using the taxonomy of Christian et al., 2010). Such additional research may help to further broaden and enlarge the knowledge that our study provided, because our analyses included only four different SJTs.

Second, in line with previously published research (Weekley, Ployhart, & Harold, 2004; Whetzel et al., 2008), females slightly outperformed males in the management sample. While there was no significant sex difference in the customer service sample, males scored higher in the administrative and graduate sample. Taking the same caveats as for the interpretation of Black–White differences across samples into account, the sex differences across samples may be due to varying personality saturation of the four SJTs (see also Whetzel et al., 2008). For example, as compared to the other SJTs, the management SJT may involve more items that target interpersonal situations. Other research revealed that females score higher on “interpersonal” traits such as warmth, affiliation, and sensitivity (see Sackett & Wilk, 1994), or agreeableness (Costa, Terracciano, & McCrae, 2001; Feingold, 1994).

This study is not without limitations. First, our samples did not include a sufficient amount of applicants with a Latin American heritage. Thus, we were unable to add to the knowledge of subgroup differences involving Hispanics. We therefore call for further studies that explore SJT subgroup differences in applicant samples to contribute to our knowledge about this ethnic group.

Second, although we drew from applicant samples, there might still be a minor range restriction in our samples. Specifically, an estimated amount of maximum 5% of applicants were screened out via a basic background check because they failed to meet the minimal qualifications (e.g., minimum level of education) stated in the job advertisement. One implication is that our estimates of *d* are somewhat conservative. Nonetheless, it is to be expected that our applicant samples are prone to much less range restriction than past studies that mainly incorporated incumbents.

Third, although we investigated subgroup differences in four distinct samples, all samples completed SJTs of the same format (i.e., SJTs with knowledge-based response instructions, a rating response format, and the consensus weighting method to set the scoring key). However, past research demonstrated that design variations in SJTs appear to influence subgroup differences. For example, knowledge-based response instructions showed slightly higher ethnic and sex group differences (Whetzel et al., 2008) which seems to be due to the higher cognitive load of knowledge-based response instructions (Whetzel et al., 2008; see also McDaniel et al., 2007) (for other examples, see Arthur et al., 2014; McDaniel, Psotka, Legree, Yost, & Weekley, 2011; Weng, Yang, Lievens, & McDaniel, 2018).

Fourth, internal consistency reliability for the SJT in the administrative sample was rather low. Low reliability might limit the accuracy of calculated subgroup differences or it might indicate more multidimensionality in the factor structure of the SJT. However, the internal consistency reliability for this SJT (0.54) was comparable to the meta-analytically estimate of internal consistency reliability for SJT scores of 0.57 reported by Campion, Ployhart, and MacKenzie (2014). In addition, internal consistency reliabilities might not reflect the most accurate estimate of reliability for SJTs that are often construct heterogeneous at the item level (e.g., Motowidlo, Crook, Kell, & Naemi, 2009; Schmitt & Chan, 2006; Whetzel & McDaniel, 2009; see also Campion et al., 2014; Sorrel et al., 2016).

Fifth, our analyses focus on subgroup differences in SJTs and do not shed a light on other aspects that influence bias against subgroups in selection (see, for overviews, Berry, 2015; Hough et al., 2001; Ployhart & Holtz, 2008). For example, we could not investigate whether the SJTs in our study show substantially different correlations to outcomes such as job performance for different ethnic or sex groups because criterion data were unavailable.

We further suggest the following directions for future research. As SJTs show subgroup differences, we should continue searching for strategies to reduce these differences (Lievens & Sackett, 2017). Research on SJTs has already demonstrated that the cognitive load of SJTs moderates the extent of subgroup differences (Dahlke & Sackett, 2017; Whetzel et al., 2008).

As one promising approach, cognitive load might be reduced by presenting the item stems via videos and by recording and rating actual behavior of the participants via a webcam. Such a constructed response multimedia test does not seem to correlate significantly with cognitive ability (De Soete, Lievens, Oostrom, & Westerveld, 2013; Lievens, De Corte, & Westerveld, 2015; Oostrom, Born, Serlie, & van der Molen, 2011) and produces small to moderate subgroup differences (Cucina, Su, Busciglio, Harris Thomas, & Thompson Peyton, 2015; De Soete et al., 2013; Lievens, Sackett, Dahlke, Oostrom, & De Soete, 2019; Lievens et al., 2015; Oostrom, Born, Serlie, & van der Molen, 2010, 2011). In a recent study, Lievens et al. (2019) empirically demonstrated that reduced majority–minority subgroup differences in constructed response multimedia tests could be attributed to unintended cognitive load in tests that aim to assess behavior related to the interpersonal domain. Constructed response multimedia tests might require less cognitive resources to perform well because they allow to respond if at least the core of the situation is understood (Hakel, 1998; Ryan & Greguras, 1998). In contrast, closed-ended response formats appear to involve higher cognitive demands because they require to evaluate the appropriateness of all given response options on a high level of detail or to detect differences between response options and to make comparative judgments between them (Marentette, Meyers, Hurtz, & Kuang, 2012). However, multimedia presentations do not always change *d* values when cognitive saturation of the test itself is high (Roth, Buster, & Bobko, 2011). Therefore, researchers and practitioners should further scrutinize the potential of coupling a constructed response format to a multimedia presentation of the SJT.

7 | CONCLUSION

This large-scale study suggests that applicant-level ethnic group differences are substantially higher than previous estimates based on incumbent data or on smaller data bases of applicant data. Hence, it may not be unexpected that SJT ethnic group differences could be in the range of d = 0.6–0.7 for applicant samples that do not suffer from much range restriction. We also found relatively small sex differences for job applicants. This affirms the guidance of Whetzel et al. (2008) that sex differences may not be associated with many cases of adverse impact. Overall, our results help researchers and decision-makers understand the actual level of subgroup differences of SJTs so they can plan selection systems and understand actual adverse impact potential.

ORCID

Christoph N. Herde https://orcid.org/0000-0002-1148-1706

ENDNOTES

[1] Note that there might be several other systematic differences between applicant and incumbent samples that might influence subgroup differences, including faking (e.g., Hough & Ones, 2001) or the possibility to retake a test (e.g., Van Iddekinge, Morgeson, Schleicher, & Campion, 2011). See the appendix from Bobko and Roth (2013) for an overview.

[2] Our sample size relating to Hispanics was too small to allow for meaningful subgroup comparisons related to this group.

[3] Although no data about the number of applicants removed due to the basic background checks were provided by selecting organizations, experience from past client projects of the business consultancy that developed the SJTs led to an estimated maximum of about 5%, suggesting little restriction.

[4] We re-ran all analyses while controlling for age. Across samples, subgroup differences did not change substantially ($\Delta\ d \le 0.03$) with the exception of Black–White differences in the graduates sample which were lower when we controlled for age (d = 0.68).

[5] A reviewer asked us to present confidence intervals across our four samples. To address this issue, we performed several small bare bones meta-analyses. For Black–White subgroup differences, we found a mean d of 0.66 (k = 4, N = 26,704) with an 80% credibility interval of 0.51–0.80 and a 95% confidence interval of 0.55–0.77 (4.88% of variance due to sampling error). For White–Asian subgroup differences, we found a mean d of 0.55 (k = 4, N = 29,131) with an 80% credibility interval from 0.48 to 0.61 and a 95% confidence interval of 0.50–0.60 (18.86% of variance due to sampling error). For sex differences, we found a mean d of 0.01 (k = 4, N = 34,209) with an 80% credibility interval from −0.11 to 0.12 and a 95% confidence interval of −0.08–0.10 (5.5% of variance due to sampling error). We caution that our confidence intervals are somewhat wide due to a relatively small k by meta-analytic standards.

REFERENCES

Arthur, W., Edwards, B. D., & Barrett, G. V. (2002). Multiple-choice and constructed response tests of ability: Race-based subgroup performance differences on alternative paper-and-pencil test formats. *Personnel Psychology*, *55*, 985–1008. https://doi.org/10.1111/j.1744-6570.2002.tb00138.x

Arthur, W., Glaze, R. M., Jarrett, S. M., White, C. D., Schurig, I., & Taylor, J. E. (2014). Comparative evaluation of three situational judgment test response formats in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. *Journal of Applied Psychology*, *99*, 535–545. https://doi.org/10.1037/a0035788

Berry, C. M. (2015). Differential validity and differential prediction of cognitive ability tests: Understanding test bias in the employment context. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*, 435–463. https://doi.org/10.1146/annurev-orgpsych-032414-111256

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

Campion, M. C., Ployhart, R. E., & MacKenzie, W. (2014). The state of research on situational judgement tests: A content analysis and directions for future research. *Human Performance*, *27*, 283–310. https://doi.org/10.1080/08959285.2014.929693

Cascio, W. F., & Aguinis, H. (2008). Staffing twenty-first-century organizations. *The Academy of Management Annals*, *2*, 133–165. https://doi.org/10.1080/19416520802211461

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

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, *63*, 83–117. https://doi.org/10.1111/j.1744-6570.2009.01163.x

Costa, Jr., P., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology*, *81*, 322–331. https://doi.org/10.1037//0022-3514.81.2.322

Cucina, J. M., Su, C., Busciglio, H. H., Harris Thomas, P., & Thompson Peyton, S. (2015). Video-based testing: A high-fidelity job simulation that demonstrates reliability, validity, and utility. *International Journal of Selection and Assessment*, *23*, 197–209. https://doi.org/10.1111/ijsa.12108

Dahlke, J. A., & Sackett, P. R. (2017). The relationship between cognitive-ability saturation and subgroup mean differences across predictors of job performance. *Journal of Applied Psychology*, *102*, 1403–1420. https://doi.org/10.1037/apl0000234

De Soete, B., Lievens, F., Oostrom, J., & Westerveld, L. (2013). Alternative predictors for dealing with the diversity–validity dilemma in personnel selection: The constructed response multimedia test. *International Journal of Selection and Assessment*, *21*, 239–250. https://doi.org/10.1111/ijsa.12034

Dean, M. A., Roth, P. L., & Bobko, P. (2008). Ethnic and gender subgroup differences in assessment center ratings: A meta-analysis. *Journal of Applied Psychology*, *93*, 685–691. https://doi.org/10.1037/0021-9010.93.3.685

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin*, *116*, 429–456. https://doi.org/10.1037/0033-2909.116.3.429

Hakel, M. D. (1998). *Beyond multiple choice: Evaluating alternatives to traditional testing for selection*. Mahwah, NJ: Erlbaum.

Hough, L. M., & Ones, D. (2001). The structure, measurement, validity, and use of personality variables in industrial, work, and organizational psychology. In N. Anderson, D. Ones, H. Sinangil, & C. Viswesvaran

(Eds.), *Handbook of industrial, work, and organizational psychology* (Vol. 1, pp. 233–277). London: Sage.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment*, *9*, 152–194. https://doi.org/10.1111/1468-2389.00171

Huffcutt, A. I., & Roth, P. L. (1998). Racial group differences in employment interview evaluations. *Journal of Applied Psychology*, *83*, 179–189. https://doi.org/10.1037/0021-9010.83.2.179

Kanning, U. P., Grewe, K., Hollenberg, S., & Hadouch, M. (2006). From the subjects' point of view. *European Journal of Psychological Assessment*, *22*, 168–176. https://doi.org/10.1027/1015-5759.22.3.168

Lievens, F. (2017). Assessing personality–situation interplay in personnel selection: Toward more integration into personality research. *European Journal of Personality*, *31*, 424–440. https://doi.org/10.1002/per.2111

Lievens, F., De Corte, W., & Westerveld, L. (2015). Understanding the building blocks of selection procedures: Effects of response fidelity on performance and validity. *Journal of Management*, *41*, 1604–1627. https://doi.org/10.1177/0149206312463941

Lievens, F., & De Soete, B. (2012). Simulations. In N. Schmitt (Ed.), *The Oxford handbook of personnel assessment and selection* (pp. 383–410). Oxford, UK: Oxford University Press.

Lievens, F., & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology*, *9*, 3–22. https://doi.org/10.1017/iop.2015.71

Lievens, F., & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. *Journal of Applied Psychology*, *96*, 927–940. https://doi.org/10.1037/a0023496

Lievens, F., & Sackett, P. R. (2017). The effects of predictor method factors on selection outcomes: A modular approach to personnel selection procedures. *Journal of Applied Psychology*, *102*, 43–66. https://doi.org/10.1037/apl0000160

Lievens, F., Sackett, P. R., & Buyse, T. (2009). The effects of response instructions on situational judgment test performance and validity in a high-stakes context. *Journal of Applied Psychology*, *94*, 1095–1101. https://doi.org/10.1037/a0014628

Lievens, F., Sackett, P. R., Dahlke, J. A., Oostrom, J. K., & De Soete, B. (2019). Constructed response formats and their effects on minority–majority differences and validity. *Journal of Applied Psychology*, *104*, 715–726. https://doi.org/10.1037/apl0000367

Lubinski, D., & Dawis, R. V. (1992). Aptitudes, skills and proficiencies. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 3, 2nd ed., pp. 1–59). Palo Alto, CA: Consulting Psychologists Press.

Lynn, R., & Irwing, P. (2004). Sex differences on the progressive matrices: A meta-analysis. *Intelligence*, *32*, 481–498. https://doi.org/10.1016/j.intell.2004.06.008

Marentette, B. J., Meyers, L. S., Hurtz, G. M., & Kuang, D. C. (2012). Order effects on situational judgment test items: A case of construct-irrelevant difficulty. *International Journal of Selection and Assessment*, *20*, 319–332. https://doi.org/10.1111/j.1468-2389.2012.00603.x

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

McDaniel, M. A., Psotka, J., Legree, P. J., Yost, A. P., & Weekley, J. A. (2011). Toward an understanding of situational judgment item validity and group differences. *Journal of Applied Psychology*, *96*, 327–336. https://doi.org/10.1037/a0021983

Motowidlo, S. J., & Beier, M. E. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. *Journal of Applied Psychology*, *95*, 321–333. https://doi.org/10.1037/a0017975

Motowidlo, S. J., Crook, A. E., Kell, H. J., & Naemi, B. (2009). Measuring procedural knowledge more simply with a single-response situational judgment test. *Journal of Business and Psychology*, *24*, 281–288. https://doi.org/10.1007/s10869-009-9106-4

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology*, *75*, 640–647. https://doi.org/10.1037/0021-9010.75.6.640

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology*, *91*, 749–761. https://doi.org/10.1037/0021-9010.91.4.749

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

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

Oostrom, J. K., Born, M. P., Serlie, A. W., & van der Molen, H. T. (2011). A multimedia situational test with a constructed-response format: Its relationship with personality, cognitive ability, job experience, and academic performance. *Journal of Personnel Psychology*, *10*, 78–88. https://doi.org/10.1027/1866-5888/a000035

Ployhart, R. E., & Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology*, *61*, 153–172. https://doi.org/10.1111/j.1744-6570.2008.00109.x

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology*, *54*, 297–330. https://doi.org/10.1111/j.1744-6570.2001.tb00094.x

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., Bobko, P., McFarland, L., & Buster, M. (2008). Work sample tests in personnel selection: A meta-analysis of black-white differences in overall and exercise scores. *Personnel Psychology*, *61*, 637–661. https://doi.org/10.1111/j.1744-6570.2008.00125.x

Roth, P. L., Bobko, P., Switzer III, F. S., & Dean, M. A. (2001). Prior selection causes biased estimates of standardized ethnic group differences: Simulation and analysis. *Personnel Psychology*, *54*, 591–617. https://doi.org/10.1111/j.1744-6570.2001.tb00224.x

Roth, P. L., Buster, M. A., & Bobko, P. (2011). Updating the trainability tests literature on Black-White subgroup differences and reconsidering criterion-related validity. *Journal of Applied Psychology*, *96*, 34–45. https://doi.org/10.1037/a0020923

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

Roth, P. L., Van Iddekinge, C. H., DeOrtentiis, P. S., Hackney, K. J., Zhang, L., & Buster, M. A. (2017). Hispanic and Asian performance on selection procedures: A narrative and meta-analytic review of 12 common predictors. *Journal of Applied Psychology*, *102*, 1178–1202. https://doi.org/10.1037/apl0000195

Ryan, A. M., & Greguras, G. J. (1998). Life is not multiple choice: Reactions to the alternatives. In M. D. Hakel (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (pp. 183–202). Mahwah, NJ: Erlbaum.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education. *American Psychologist*, *56*, 302–318. https://doi.org/10.1037/0003-066X.56.4.302

Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist*, *49*, 929–954. https://doi.org/10.1037/0003-066X.49.11.929

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

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods*, *19*, 506–532. https://doi.org/10.1177/1094428116630065

Van Iddekinge, C., Morgeson, F., Schleicher, D., & Campion, M. (2011). Can I retake it? Exploring subgroup differences and criterion-related validity in promotion retesting. *Journal of Applied Psychology*, *96*, 941–955. https://doi.org/10.1037/a0023562

Weekley, J. A., Ployhart, R. E., & Harold, C. M. (2004). Personality and situational judgment tests across applicant and incumbent settings: An examination of validity, measurement, and subgroup differences. *Human Performance*, *17*, 433–461. https://doi.org/10.1207/s15327043hup1704_5

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

Weng, Q. D., Yang, H., Lievens, F., & McDaniel, M. A. (2018). Optimizing the validity of situational judgment tests: The importance of scoring methods. *Journal of Vocational Behavior*, *104*, 199–209. https://doi.org/10.1016/j.jvb.2017.11.005

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

How to cite this article: Herde CN, Lievens F, Jackson DJR, Shalfrooshan A, Roth PL. Subgroup differences in situational judgment test scores: Evidence from large applicant samples. *Int J Select Assess*. 2020;28:45–54. https://doi.org/10.1111/ijsa.12269

APPENDIX

TABLE A1 Overview of SJTs by sample

Scenarios	Example
	Management sample
24	You manage a department. One of the most experienced members of the department has considerable knowledge about processes and historical issues within the organization. His knowledge has been valuable in supporting the less experienced team members in their tasks. Up until recently he has been an effective member of your team with a good attendance record. However, recently he has taken sick leave on several occasions. Three members of the department have also raised concerns about him, as he has been unhelpful and unpleasant to them *Rate the effectiveness of the following actions* a. Meet with the team member and ask him how his work is going generally and whether he has any issues that he would like to discuss with you b. Tell the team member that his recent performance and attendance has been unacceptable and you expect to see an immediate and substantial improvement in both c. Take no action for now and wait to see whether the issues resolve themselves as, until recently, he has been a good performer and the issues may only be temporary d. Meet with the team member to discuss the concerns raised and establish whether there are any underlying issues, and then agree specific actions to address them
	Graduate sample
20	You have been asked to provide some data to a Senior Manager. She has demanded the data by the end of today for an important meeting tomorrow. The data is complex and if you conduct a thorough check to make sure it is all correct, this will take you until the end of tomorrow. No-one else is available to help you. It is unlikely that anyone would notice an error in the data, but it will be used to make decisions *Rate the effectiveness of the following actions* a. Explain to the Senior Manager that you can provide the data by the end of today, but this will not give you enough time to check it thoroughly so there may be errors b. Ask the Senior Manager what parts of the data are most important for her meeting tomorrow and then focus on checking these parts of the data before the end of today c. Check what you can before the end of today, then pass it to the Senior Manager and inform her you will continue checking it tomorrow and let her know if you find any additional errors d. Explain to the Senior Manager that it is not possible for you to provide the data before the end of tomorrow
	Customer service sample
20	You are working in a branch of a busy high street bank. The bank offers a large range of financial products to customers including insurance, savings, current accounts, credit cards and mortgages. You are working on the customer service desk when a customer approaches you. She explains that she is having difficulty completing an application form for a new current account. You remember explaining to the same customer yesterday how to complete the form, but she is still having problems. The instructions for completing the form can also be found on the bank's website *Rate the effectiveness of the following actions* a. Offer to sit down and help the customer to complete the application form, making sure she fully understands the information required in each section b. Ask a colleague to help the customer complete the form, as you have already tried to explain it to her once and she did not understand c. Suggest that the customer takes the form away to complete at home and has a look at the bank's website to find out how to complete it d. Suggest that the customer fills out the form as best she can and tell her that any mistakes should be picked up when the form is processed
	Administrative sample
20	Your Manager has asked you to format and print a financial document which will be collected by a courier in an hour's time. Your Manager is now in an important meeting and is unavailable until this evening. Whilst formatting the document, you notice that some of the data does not appear to match up. You are unsure if this is deliberate *Rate the effectiveness of the following actions* a. Concentrate on formatting and printing the document in time for the courier to collect it, but mention the data issue to your Manager when she is next available b. Try to find a colleague who knows about this financial document and ask them to check quickly whether the data is correct c. Rearrange the courier for tomorrow and send an email to your Manager suggesting she checks the report before you send the final copy d. Quickly make the changes you think are necessary to the data, and format and print the document ready for the courier