5-2019

Constructed response formats and their effects on minority-majority differences and validity

Filip LIEVENS
Singapore Management University, filiplievens@smu.edu.sg

Paul R. SACKETT
University of Minnesota - Twin Cities

Jeffrey DAHLKE
University of Minnesota - Twin Cities

Janneke OOSTROM
Vrije University

Britt DE SOETE
Ghent University

DOI: https://doi.org/10.1037/apl0000367

Citation

LIEVENS, Filip; SACKETT, Paul R.; DAHLKE, Jeffrey; OOSTROM, Janneke; and DE SOETE, Britt. Constructed response formats and their effects on minority-majority differences and validity. (2019). *Journal of Applied Psychology*. 104, (5), 715-726. Research Collection Lee Kong Chian School Of Business.
Available at: https://ink.library.smu.edu.sg/lkcsb_research/5953

Constructed Response Formats and Their Effects on Minority-Majority Differences and Validity

Filip Lievens

Singapore Management University

Paul R. Sackett & Jeffrey Dahlke

University of Minnesota-Twin Cities

Janneke K. Oostrom

Vrije Universiteit Amsterdam

Britt De Soete

Cubiks Belgium

In press *Journal of Applied Psychology*

Author note

Filip Lievens, Lee Kong Chian School of Business, Singapore Management University, Singapore. Paul R. Sackett, Department of Psychology, University of Minnesota-Twin Cities. Jeffrey Dahlke, Department of Psychology, University of Minnesota-Twin Cities. Janneke K. Oostrom, Department of Management and Organization, School of Business and Economics, Vrije Universiteit Amsterdam. Britt De Soete, Cubiks Belgium.

Correspondence concerning this article should be addressed to Filip Lievens, Lee Kong Chian School of Business, Singapore Management University, 50 Stamford Road, Singapore 178899. E-mail: filiplievens@smu.edu.sg.

Part of the funding for the studies reported came from a PhD fellowship of the Research Foundation – Flanders (1162012N) and from the SIOP Foundation via the 2014 Adverse Impact Reduction Research Initiative and Action (AIRRIA) grant. We are indebted to Anne Marie Ryan for her excellent suggestions on an earlier draft of this paper. We also thank Christoph N. Herde and Pat Sackett for their editorial assistance.

Abstract

The inflow of immigrants challenges organizations to consider alternative selection procedures that reduce potential minority (immigrants)-majority (natives) differences, while maintaining valid predictions of performance. To deal with this challenge, this paper proposes response format as a practically and theoretically relevant factor for Situational Judgment Tests (SJTs). We examine a range of response format categories (from traditional multiple choice formats to more innovative constructed response formats) and conceptually link these response formats to mechanisms underlying minority-majority differences. Two field experiments are conducted with SJTs. Study 1 (274 job seekers) contrasts minority-majority differences in scores on a multiple choice vs. a written constructed response format. Written constructed responses produce much smaller minority-majority differences (*d* = .28 vs. *d* = .92). In Study 2 (269 incumbents), scores on a written constructed vs. an audiovisual constructed format are compared. The audiovisual format further reduces minority-majority differences (*d* = .09 vs. *d* = .41), with validities remaining the same. Results are suggestive of cognitive load as a contributor to the reduction in minority-majority differences, as are rater effects: Scores of raters evaluating transcribed audiovisual responses, which anonymized test-takers, produce larger differences. In sum, altering response modality via more realistic response formats (i.e., the audiovisual constructed format) leads to significant reductions in minority-majority differences without impairing criterion-related validity. Implications for selection theory and practice are discussed.

Keywords: Minorities, immigrants, response format, subgroup differences, validity, Situational Judgment Tests

Running Head: SJT RESPONSE FORMAT AND MINORITY-MAJORITY DIFFERENCES

Constructed Response Formats and Their Effects on Minority-Majority Differences and Validity

According to the International Organization for Migration, worldwide immigration is increasing exponentially with an expected 405 million migrants by 2050 (IOM, 2018). Hence, immigrants are an increasingly large and important group in applicant pools of many Western nations (Binggeli, Dietz, & Krings, 2013; Harrison, Harrison, & Shafffer, 2018). This immigrant inflow challenges organizations to consider selection procedures that reduce differences between this new minority group and the majority population, while still providing valid predictions.

Situational Judgment Tests (SJTs) are often proposed to be such selection procedures (Landy, 2007), even though an issue is that minority-majority differences in SJT scores still vary substantially. The most recent large-scale review found *d* values between .19 and 1.02 with an average of. 38 (Bobko & Roth, 2013). One of the reasons for minority-majority differences in SJT scores is that in SJTs a multiple-choice (MC) response format is typically used. Although such an MC format permits standardization and rapid scoring, it might also add incidental cognitive load to a primarily interpersonal measure like an SJT. Cognitive load means that SJT scores correlate with cognitive ability and thus also capture cognitive variance. Given that reducing incidental cognitive load in mostly interpersonal measures like SJTs is key for reducing minority-majority differences (Dahlke & Sackett, 2017), response formats other than MC should be scrutinized (see Table 1 for a framework of response formats). Examples are *written constructed* (e.g., producing a response in writing) and *audiovisual constructed response formats* (e.g., responding orally via a webcam).

Although these newer constructed response (CR) formats have become popular due to technological advances (text analytics and webcams), research on them is in its infancy and an “area where practice may have moved too quickly ahead of science” (Cucina et al., 2015, p.197). Prior studies (see Table 1) mainly established the validity of these newer CR SJT formats, whereas the effects of SJT response format on minority-majority differences remained unexplored (Bobko & Roth, 2013; Edwards & Arthur, 2007). So, these studies did not address the key question whether such newer SJT response formats indeed reduce minority-majority differences and which factors might cause this.

This study contributes to response format and SJT research by (1) examining a range of response formats in terms of their effects on minority-majority differences and validity and (2) proposing and testing theoretical rationales underlying their effects (cognitive load, divergent thinking, test motivation, rater effects). Study 1 compares immigrants' vs. natives' scores on an MC vs. written CR SJT. Study 2 goes one step further by comparing written CR to audiovisual CR SJT responses in terms of immigrant-native differences and validity for predicting interpersonal criteria.

HYPOTHESES

To better understand test score differences between different subgroups (e.g., Blacks-Whites; immigrants-natives), Newman and colleagues (Cottrell, Newman, & Roisman, 2015; Outtz & Newman, 2009; see also Rindermann, Becker, & Coyle, 2016) developed a comprehensive model that distinguished between genetic, environmental (e.g., socio-economic, educational, cultural, family environment), and test-related factors. Given this study's aims, we focus on test-related factors, while trying to control for some of the other factors. A central tenet of the model is that a test score should not contain performance-irrelevant ethnicity-related variance. In the context of interpersonal measurement via SJTs, we posit that an MC format contains at least three possible sources of performance-irrelevant ethnicity-related variance.

The cognitive loading of a test (i.e., the correlation of a test score with general cognitive ability) is a first factor because there exists substantial evidence of immigrant-native differences on cognitive ability tests. Robie et al. (2017) compared immigrant-native differences across 29 countries and found a *d* of .53 in favor of the majority group; a meta-analysis by Te Nijenhuis, Willigers, Dragt, and Van der Flier (2016) reported similar results. Moreover, this meta-analysis revealed that immigrant-native differences become larger when test scores had a higher cognitive loading and that this test-related factor was more important than language/cultural factors. This is a pivotal result because it highlights the importance of avoiding unintended cognitive load in primarily interpersonal measures. If one uses an MC format in SJTs and if the MC format has a higher cognitive load than a written CR format, then this might lead to immigrant-native differences on SJT scores.

There are indeed reasons why MC formats might be expected to have greater cognitive load: An MC format requires reading all response options, detecting nuances among them, and making a comparative judgment to select the best one (Marentette, Meyers, Hurtz, & Kuang, 2012). In contrast, CR formats demand fewer cognitive resources because they permit responding when the core message is understood (Hakel, 1998; Ryan & Greguras, 1998). So, we hypothesize smaller minority-majority differences for SJT scores in a CR format than in an MC format (H1) and that the reduced cognitive load of written CR SJT scores partially accounts for smaller minority-majority differences in these SJT scores as compared to the MC scores. Formally, we posit moderated mediation: cognitive ability mediates the ethnicity-SJT relationship, with format (MC vs. CR) moderating the cognitive ability – SJT relation (H2a).

Apart from cognitive load, a second test-related factor is that MC responses typically rely on convergent thinking (i.e., producing or identifying the single correct answer to a question or problem, Cropley, 2006; Outtz, Goldstein, & Ferreter, 2006; Outtz & Newman, 2009). Conversely, CR modalities also allow for divergent thinking (i.e., generating multiple solutions to a question or using multiple ways to achieve a solution, Guilford, 1950). So, we expect SJT scores in the CR format to be more saturated with divergent thinking than those in the MC format. This difference between MC and written CR formats in terms of divergent thinking is important for reducing minority-majority differences only when immigrants score higher on divergent thinking. Indeed, creativity seems to be higher among immigrants (Simonton, 1999). Research also confirmed that living in another country is related to higher divergent thinking, independently from factors such as openness or cognitive ability (Maddux & Galinsky, 2009). According to Leung, Maddux, Galinsky, and Chiu (2008, p.5), multicultural experience fosters creativity by shaking up knowledge, providing novel ideas, bringing diverse perspectives, and encouraging the combination of dissimilar ideas (see also Çelik, Storme, & Forthmann, 2016; Goclowska & Crisp, 2014). So, we expect the increased divergent thinking saturation of SJT scores in the written CR format to partially account for smaller minority-majority differences in SJT scores as compared to the MC format. In moderated mediation terms: divergent thinking mediates the ethnicity-SJT relationship, with format (MC vs. CR) moderating the divergent thinking - SJT relation (H2b).

Finally, test-taking motivation represents a third potential test-related factor underlying reductions in minority-majority differences in written CR formats. Ethnicity has been found to play a modest but consistent role in test motivation and perceptions towards tests with an MC format: Minorities display lower test motivation and less favorable test perceptions, which then translates into lower test performance (Chan, Schmitt, DeShon, Clause, & Delbridge, 1997; Ryan, 2001; Schmit & Ryan, 1992). Although these results were obtained in the US, recent research on immigrant samples in Europe has confirmed these findings of minorities displaying lower test motivation and less favorable test perceptions than the majority group (Oostrom & De Soete, 2016). Importantly, these results relate primarily to tests with an MC format. The situation might be different for alternative formats. Indeed, in Europe there is evidence that immigrants prefer such alternative response formats because they might have experienced a history of lack of success with traditional MC formats and might have developed negative attitudes towards them (Hiemstra, Derous, Serlie, & Van der Molen, 2012). Similarly, in the US, minorities showed more favorable attitudes towards CR formats, which reduced Black-White differences on a knowledge test (Edwards & Arthur, 2007). So, when a CR format is used, immigrants' higher test motivation might raise their SJT performance, thereby reducing minority-majority differences in SJT scores. So, we expect minorities' higher test motivation to an SJT in written CR format to partially account for the smaller minority-majority differences in SJT scores in this format as compared to an MC format. In moderated mediation terms: motivation mediates the ethnicity-SJT relationship, with format (MC vs. CR) moderating the motivation – SJT relation (H2c).

STUDY 1

Sample

The sample consisted of 274 job seekers (97 men; 177 women) registered at the same Belgian employment agency. Participants' mean age was 31.95 (*SD* = 10.02). There were 186 majority (natives) and 88 ethnic minority members (immigrants). In line with the official operationalization used across Western Europe, ethnic minority members were people with at least one biological parent or two or more biological grandparents originating from non-Western-European/Northern-American countries (VESOC, 2003).

Of the ethnic minorities, 28% had a college/university degree (vs. 42% for native job seekers). Ethnic minorities averaged 5.3 yrs. of job experience (vs. 9.6 yrs. for natives), had different nationalities (from Middle East, Africa, or Eastern Europe), and only 29% of them spoke Dutch in the home. This matches the ethnic minority profile in the EU (less education and job experience, and speaking another home language (Te Nijenhuis et al., 2016).

Procedure and Design

Participants were told by the employment agency that they would be completing several tests for research purposes and that they would obtain detailed feedback and a € 5,00 store coupon. After being e-mailed the testing website link, they were instructed to complete the study materials in an environment without distraction.

A between-subjects design was used in which respondents were randomly assigned to one of two response formats: Participants in the MC condition (N = 137) received the SJT with five potential response options and had to select the most effective one. In the written CR condition (N = 134), people were presented with the same SJT but had to type the most effective response in a text box.

Measures

Situational Judgment Test. We used the Work Judgment Survey (Smith & McDaniel, 1998) which consists of 31 items about work situations describing coworker and employee-supervisor problems. Appendix A explains how we aligned the scoring in the MC and CR versions. For analysis purposes, we standardized SJT scores within conditions.

Cognitive ability. The cognitive ability test consisted of a total of 24 items (divided into three subtests: verbal, numeric, or abstract) followed by five response options. This test measures general cognitive ability. A time constraint of 22 minutes was imposed. Subtests correlated .54 to .59. A composite ability score was thus computed. Prior research (Minnaert, 1996) found evidence for its adequate reliability (.84) and validity for predicting GPA (.36).

Test motivation. After the SJT, test motivation was measured with three items from Arvey, Strickland, Drauden, and Martin (1990) via a 5-point Likert scale (α = .82). An example item was “I was very motivated to perform well on this test”.

Divergent thinking. We measured divergent thinking via the Unusual Uses Test (Iscoe & Pierce-Jones, 1964; see also Çelik et al., 2016), which is a common creativity task in which respondents invent as many different uses as possible for a common object. The test consisted of 10 items (α = .92). An example item is ‘Come up with as many different uses for the object *newspaper’*. Examples of potential answers are *to read*, *to fold a boat*, *as a hat*, etc. There was no time constraint. A divergent thinking score was calculated on the basis of the mean of the total number of separate functional categories given by a respondent per item.

Other measures. Home language spoken in the home was coded as 0 if participants self-reported Dutch as the primary home language, and coded as 1 otherwise. In addition, participants reported their job experience in years.

Results and Discussion

We started with multiple group measurement invariance analyses (using Mplus). Scores on the 31 SJT items across the two response formats served as indicator variables. Prior to testing measurement invariance, we sought to establish a baseline model. Consistent with prior SJT research (e.g., Krumm et al., 2015), the models specified generally showed a poor model fit. A model with all items loading on one factor showed the best fit. Two items with negative factor loadings had to be removed so that the total number of items was 29 in all our analyses. After establishing the one-factor model as the baseline model, we continued with an increasingly restrictive series of measurement invariance tests. As shown in Appendix B, there was evidence of form and metric invariance for this model. So, the measurement structure underlying scores on the SJT was invariant across the two response formats.

To test H1 (subgroup differences are smaller for CR scores than for MC scores) we computed the standardized mean difference (d) on the SJT by ethnicity for each format. H1 was confirmed, with $d = .92$ for MC scores and $d = .28$ for CR scores (both in favor of the majority group). The difference is statistically significant ($p < .05$). At a practical level, using a CR format instead of an MC format led to a 70% reduction in subgroup differences on an

SJT. So, as a key conclusion of Study 1, a written CR format displayed significantly smaller minority-majority differences in performance as compared to an MC format.

Table 2 and Table 3 show correlations among the study variables in the two conditions. The ability- SJT MC format correlation (.47) was significantly ($p < .01$) higher than the ability- SJT CR correlation (.14), suggesting lower cognitive load for the latter[1]. To test the hypotheses that cognitive ability (H2a), divergent thinking (H2b), and motivation (H2c) contribute to the smaller majority-minority subgroup difference for the CR format relative to the MC one, we used Hayes's (2018) SPSS PROCESS macro to test a moderated mediation model with the above three variables as mediators of the ethnicity –SJT relationship, and with response format as a moderator of the relationship between the three mediators and SJT scores. Figure 1 shows this model's results. Language spoken in the home and job experience were included as control variables, so all effects are net of these controls.

Figure 1 shows that ethnicity is related to cognitive ability, and the ability-SJT relationship is moderated by format. Examining conditional effects, and using bootstrapped standard errors, the ability-SJT relationship is .42 (*SE* = .10) for MC and .03 (*SE* = .08) for CR. The indirect effect of ethnicity on SJT is -.26 (*SE* = .09) for MC and -.02 (*SE* = .06) for CR. Thus, ability mediates the ethnicity-SJT relationship for MC, and not for CR, providing support for H2a. In contrast, the hypothesized moderated mediation effects were not found for divergent thinking (H2b) and motivation (H2c). While ethnicity was significantly related to divergent thinking (-.49), format did not moderate the divergent thinking-SJT or motivation-SJT relationships, nor was divergent thinking or motivation found to mediate the ethnicity-SJT relationships for either MC or CR. In sum, cognitive ability emerged as the central variable mediating the ethnicity-SJT relationship, with ability showing a strong relationship with MC SJT scores and a negligible relationship with CR SJT scores.

[1] To rule out that the higher cognitive load of an MC vs. constructed response SJT was due to the fact that the ability test also had an MC format we also ran analyses with a constructed response knowledge test (correlated with the ability test and administered for purposes separate from this study). Results were similar.

Importantly, the direct path from ethnicity to SJT was not significant, further highlighting the mediating role of cognitive ability.

Although Study 1 suggests the CR format in SJTs as a viable strategy for reducing subgroup differences, some questions remain unanswered. First, we do not know whether further reductions can be obtained with even higher fidelity response formats (i.e., audiovisual CRs; see Table 1). Second, Study 1 did not address the effects on validity for predicting interpersonal criteria.

STUDY 2

Study 2 compares two CR formats (written vs. audiovisual responses in front of a webcam) in terms of minority-majority differences as well as validity for predicting interpersonal criteria. We expect audiovisual responses to further decrease cognitive load and thus to lead to smaller SJT minority-majority differences for two reasons. First, linguistic research shows that oral responses require fewer cognitive resources than written ones (Bourdin & Fayol, 2002). When immigrants learn a second language, writing the new language is also understandably more effortful than speaking it (Van Tubergen, 2010). So, spoken responses seem to demand less cognitive effort than written ones. Second, the audiovisual format requires responding more spontaneously with limited reflection time (Lievens, De Corte, & Westerveld, 2015). In contrast, writing responses might be a more thoughtful process as one might reflect before answering, re-read, and correct responses. Thus*,* we expect smaller minority-majority differences for SJT scores from an audiovisual CR than from a written CR format (H3), and that the reduced cognitive load of SJT scores in the audiovisual condition partially accounts for smaller minority-majority differences in this condition. We frame this again as moderated mediation: cognitive ability mediates the ethnicity-SJT relationship, with format (written vs. audiovisual CR) moderating the cognitive ability – SJT relationship (H4).

Besides cognitive load, these response formats also differ in their sensitivity to rater effects: unlike the written response format, the audiovisual one reveals people's ethnic background, which might activate rater stereotypes (i.e., category-based attributes applied to a group of people as a result of beliefs and expectations about the group's members, Agars, 2004). Reliance on stereotypes may result in biased ratings (Biernat, 2003) and thus in unintended rater effects in the judgment process. Transcribing audiovisual responses[2] eliminates visual/auditory cues, and can provide insight into potential rater effects. That is: Similar subgroup differences across the audiovisual and transcribed responses argue against rater effects, while a discrepancy in subgroup differences suggests rater effects.

Stereotype rater effects may operate in two directions (Biernat, 2003; Koch, D'Mello, & Sackett, 2015). On one hand, stereotyping may lead to bias in ratings based on the ratee's group membership so that disadvantaged groups systematically receive lower ratings. Prior research showed that visual or auditory information might negatively influence ratings in the case of Muslim applicants (King & Ahmad, 2010) and applicants with a foreign accent (Hosoda & Stone-Romero, 2010; Purkiss, Perrewé, Gillespie, Mayes, & Ferris, 2006). Similar stereotypes might occur for response formats that disclose visual ethnicity information (Barron, Hebl, & King, 2011; Frazer & Wiersma, 2001; Kaiser & Pratt-Hyatt, 2009).

On the other hand, a reverse stereotyping effect might also take place. As people make within-group comparisons in their judgments, they may adopt shifting standards when evaluating members from different social groups (Biernat & Manis, 1994). If lower performance is expected for a particular ethnic group, a lower comparison standard might be used for evaluating those ethnic group members, whereas a higher comparison standard might be used for other ethnic group members. As a result, a *good* rating for one ethnic group might mean something different than a *good* rating for another group. Prior experimental research

[2] Although the transcribed format is used here for research purposes, it has also practical value as modern technology enables organizations to transcribe responses and use these as the basis for making ratings.

documented the effect of shifting standards when evaluating minority vs. majority members or males vs. females (e.g., Biernat & Sesko, 2013; Collins, Biernat, & Eidelman, 2009; see Holder & Kessels, 2017, for research in immigrant samples). Relatedly, people’s motivation to suppress prejudice might play a role (Crandall, Eshleman, & O'Brien, 2002; Devine, Plant, Amodio, Harmon-Jones, & Vance, 2002; Plant & Devine, 1998). That is, in the audiovisual format, raters may feel accountable for not being prejudiced because the information on ethnicity is available. As a result of these mechanisms (shifting standards and motivation to suppress prejudice), subgroup differences might be smaller for the audiovisual responses than for their transcriptions because the latter do not disclose respondents’ ethnicity. Given the opposing conceptual perspectives, we formulate a research question: Are minority-majority differences for audiovisual responses significantly larger than for their transcriptions?

In the presence of adverse impact the Uniform Guidelines on Employee Selection Procedures (Equal Employment Opportunity Commission, 1978) calls for a search for approaches that reduce subgroup differences, but still result in comparable validities. So, reductions in subgroup differences should not come at the expense of a validity decrease. As the audiovisual response mode requires showing actual interpersonal behavior and thus ensures high point-to-point criterion correspondence with such behavior on the job, we anticipate its scores to have adequate validity for predicting interpersonal criteria. This should compensate for the lower cognitive load of the audiovisual responses. We expect a similar compensation for the transcribed format: Although in the transcription nonverbal information that provides valid cues in interpersonal interview contexts (DeGroot & Motowidlo, 1999) is lost, there could be gains from avoiding stereotype effects when raters are blind to ethnicity. So, we hypothesize no differences in the validity of the audiovisual response format, or its transcribed version, compared to the written format (H5) for predicting interpersonal criteria. We also expect no differences in validity by subgroup (H6).

Method

Sample

The sample consisted of 269 (81 men; 188 women) government employees in Belgium. Participants' mean age was 36.55 (*SD* = 10). All participants were working in white-collar jobs of similar level that required coworker and client interactions (e.g., project-related or administrative jobs for education, transport, or other departments). There were 169 ethnic majority and 100 ethnic minority members (see Study 1 for the operationalization of ethnic minority status), with only 37% having a college/university degree (vs. 74% for natives). Ethnic minorities averaged 5.9 yrs. of job experience (vs. 12.5 yrs. for natives), had 19 different nationalities (from Middle East, Africa, or Eastern Europe), and only 27% spoke Dutch in the home.

Procedure and Design

After being invited by e-mail, test administration proceeded in separate rooms on a PC with a webcam. Participants started with the CR multimedia SJT. At a specific point in each multimedia SJT item, the scene froze and participants were required to respond as if they actually took part in the situation. This was the point where the two formats differed. In the written CR format (8 fragments), participants were presented with a text box and asked to note down their response. In the audiovisual CR format (8 other fragments), a webcam started recording their response when the video ended. We employed a within-subjects design. To control for item, order, fatigue, and practice effects, we developed four item sets so that each participant received eight items with a written CR format and eight items with an audiovisual CR format, with response mode order, item set, and item sequence being counterbalanced. Participants were randomly assigned to one of the sets. After the SJT, they completed the cognitive ability test. At the end, they received a €10,00 coupon and a feedback report.

Measures

Multimedia SJT. The multimedia SJT aimed to measure interpersonal competencies at work such as social sensitivity, working with others, listening, etc. (Klein, DeRouin, & Salas, 2006). We followed common practices for developing multimedia SJTs (e.g., Weekley & Jones, 1997): interviews to gather incidents, subject matter expert review, editing of incidents, and script development. A professional filming company was in charge of filming and postproduction. All of this resulted in an SJT with 16 video-based scenarios.

Assessors, assessor training, rating process, and ratings. Assessors received one day of training in line with the International Task Force on Assessment Center Guidelines (2009). They were taught to use behaviorally anchored rating scales for rating participants' interpersonal skills (Klein et al., 2006). A 5-point scale ranging from *1 = poor* to *5 = outstanding* was used. Video and text fragments of participant performances were also presented and evaluated. Finally, assessors individually rated 10 practice responses. Only in case of sufficient inter-rater agreement (ICC[1,2] > .70), were they certified as assessors. The final assessor pool consisted of 24 I/O psychology graduates (21 females; mean age = 23 yrs., *SD* = 3). In the actual rating sessions, audiovisual, transcribed, and written responses were randomly assigned to two assessors. In total, 12,912 ratings were made.

Cognitive ability test. The same test as in Study 1 was used but this time with 19 items and a time constraint of 20 minutes. We again computed a composite ability score.

Other variables. Language spoken in the home and job experience were measured in the same way as in Study 1.

Criterion measure. We used peer ratings because peers view employees through an interpersonal lens (Conway & Huffcutt, 1997; Viswesvaran, Ones, & Schmidt, 1996). After asking participants to nominate peers, we randomly selected two peers to provide an evaluation of the participant's interpersonal work performance on several dimensions (α = .74) via the relative percentile method (Goffin, Gellatly, Paunonen, Jackson, & Meyer, 1996).

Peers knew their ratings would be used only for research. We received responses for 193 employees (73% response rate; 123 majority members, 70 minority members). All peers had daily/weekly contact with the participant. Agreement among peer ratings equaled .56 (ICC[1,2]), which is in line with meta-analyses on peer ratings (Viswesvaran et al., 1996).

Results and Discussion

Similar to Study 1, we ran multiple group measurement invariance analyses using Mplus. Participants' scores on the eight multimedia SJT items across the three formats served as indicators. Again, the best fit was obtained for a one-factor model, even though fit indices were below acceptable levels. Appendix B shows evidence of form and metric invariance for this model across written, audiovisual, and transcribed CR scores.

In support of H3 (smaller differences for audiovisual than for written CR scores) we found *d*'s of .09 for the audiovisual format and .41 for the written CR format in favor of the majority group. This difference between the two scores is significant ($p < .01$). So, importantly, minority-majority differences were smaller for the audiovisual CR scores than for the written CR scores. So, minority members seem to benefit from a high-fidelity answering modality. Of particular interest is that ratings of transcriptions of the audiovisual scores resulted in a subgroup *d* of .30, which was not significantly different from the $d = .41$ for the written one. Likely, shifting rating standards (Biernat & Manis, 1994) and motivation to suppress prejudice (Crandall et al., 2002) account for these findings as the raters see the candidates' video and thus are aware of their ethnicity in the audiovisual CR condition but not in the transcribed one.

Table 4 shows correlations among study variables. There were correlations of .16 and .20 between ability and the audiovisual and written CR SJTs. So, written CR scores had a higher cognitive load but the difference was not significant. H4 posited cognitive ability to mediate the relationship between ethnicity and SJT performance, with format moderating the ability-SJT relationship. Thus, as in Study 1, moderated mediation is hypothesized. A different analytic approach is required, as format is a within-person variable in Study 2, while it was a between-person variable in Study 1. In Study 2, a written score, an audiovisual score, and a transcribed score are available per person. Thus, a multilevel model is needed, as format is a nested level-1 variable and ethnicity, cognitive ability, and two control variables (language in the home and job experience) were level-2 between-subjects variables. We tested the moderated mediation model using the "lavaan" R package. Figure 2 shows the model's results. All results are net of the control variables (language in the home, job experience).

Examining total effects, ethnicity is related to SJT scores only in the written SJT format (-.52). As hypothesized in H4, this effect is partially mediated by cognitive ability, as the indirect effect is significant (-.12). For the audiovisual format, the ethnicity-SJT total effect relationship is not significant, though the indirect effect mediated via cognitive ability is significant, reinforcing the recent observation that, contrary to prior thinking, a mediating relationship is not dependent on a significant total effect (Hayes, 2018). For the transcribed format, the total, direct, and indirect ethnicity- SJT relationships are all non-significant.

H5 stated that the relationship between SJT scores and the criterion is comparable for all three formats. As shown in Table 4, *r*s were between .15 and .16, in support of H5. Importantly, this shows that the reduction in minority-majority differences due to response format did not lead to lower validity. There was also support for H6 as the validity correlations did not differ by ethnicity (*r*s between .10 and .18, which were not significantly different from one another). Note that there were also no slope/intercept differences in prediction by ethnicity.

GENERAL DISCUSSION

This paper proposed response format as an as-yet unexplored driver of minority-majority differences and contributed to response format and SJT research by examining two newer written CR formats in SJTs: written and audiovisual formats. Importantly, this paper found substantial reductions in minority-majority differences for these newer CR formats. Across the studies, *d'*s were reduced from .92 (MC format) to .09 (audiovisual CR format). Importantly, the audiovisual CR format also met the provision of "less adverse impact with equal validity". Employers often search for such selection alternatives and this study shows how response format modifications constitute a viable alternative in the interpersonal domain.

Conceptually, this study contributed to theorizing about the mechanisms by which response format moderates minority-majority differences. Results across both studies were suggestive of the role of unwanted cognitive load in instruments that have a primarily interpersonal orientation. We found no evidence of test motivation or divergent thinking and some evidence of rater effects (smaller differences when raters knew people's ethnicity). Future studies are needed to examine other potential mechanisms (e.g., communication styles, Gudykunst et al., 1996; Helms, 1992). Think-aloud or cognitive interviewing approaches (Beatty & Willis, 2007; Oostrom & Born, 2014) might be used to shed light on such mechanisms. As rater effects play a role, we also need research on automated scoring techniques such as latent semantic analysis (e.g., Lenhard, Baier, Hoffmann, & Schneider, 2007) and social sensing (Schmid Mast, Gatica-Perez, Frauendorfer, Nguyen, & Choudhury, 2015). This could replace the time-intensive hand scoring of CRs. For example, in Study 2, it took raters about 35 minutes to rate a respondent's responses.

As a final contribution, this study extended subgroup differences research by focusing on a new minority group (immigrants). Due to worldwide immigration, such immigrant pools become increasingly prevalent (Harrison et al., 2018). We also went beyond examining immigrant vs. native differences on ability tests. Given that the small and diverse minority sample of both studies did not permit differentiating according to specific ethnic background[3], future studies should refine our results with larger samples and other criteria.

Future studies are also needed to examine how response format choice interacts with stimulus format choice. Although high-tech stimulus formats (3D animation, avatars) are now advocated (Fetzer & Tuzinski, 2013), they are still often coupled with MC responses. This study should also stimulate more comparative research on newer response formats. As a limitation of our studies is that they were conducted for research purposes (with either job seekers or incumbents), it is best to do such comparative research with actual applicants. Yet, such a study remains a challenge in actual selection practice because format manipulations might be seen as a differential treatment (Edwards & Arthur, 2007).

In sum, this paper is the first to provide unprecedented evidence that response format is a practically and theoretically relevant factor in the search for interpersonal selection procedures (SJTs) that reduce minority-majority differences. Although MC response formats have typically dominated assessment, this paper shows that using CR formats (and especially the audiovisual format) also benefit organizations that strive to increase diversity inflow while maintaining validity. Moreover, this actionable advice fits well in the recent deployment of technology to make CR more feasible (e.g., via text analytics and webcams). Given that we took a modular approach (Lievens & Sackett, 2017) and focused on one key component (response format) underlying a variety of selection procedures, this strategy can be adopted across various simulation-based procedures in the interpersonal domain (high-fidelity simulations such as assessment center exercises/work samples, low-fidelity simulations such as SJTs, and their hybrids), thereby fully acknowledging that other factors (e.g., time, cost) also come into play in these adoption decisions.

[3] Effect sizes were similar for the largest ethnic group (Turks/Moroccans). For instance, in Study 1, *d* was .96 for MC responses (*d* = .96 for all minorities) and *d* was .36 for written constructed ones (*d* = .30 for all minorities).

References

Agars, M. D. (2004). Reconsidering the impact of gender stereotypes on the advancement of women in organizations. *Psychology of Women Quarterly, 28*, 103-111.

Arthur, W., Jr., 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.

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

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43*, 695-716.

Barron, L. G., Hebl, M., & King, E. B. (2011). Effects of manifest ethnic identification on employment discrimination. *Cultural Diversity & Ethnic Minority Psychology, 17*, 23-30.

Beatty, P. C. & Willis G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71*, 287-311.

Biernat, M. (2003). Toward a broader view of social stereotyping. *American Psychologist, 58*, 1019-1027.

Biernat, M., & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of Personality and Social Psychology, 66*, 5-20.

Biernat, M., & Sesko, A. K. (2013). Communicating about others: Motivations and consequences of race-based impressions. *Journal of Experimental Social Psychology, 49*, 138-143.

Binggeli, S., Dietz, J., & Krings, F. (2013). Immigrants: A forgotten minority. *Industrial and Organizational Psychology-Perspectives on Science and Practice, 6*, 107-113.

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.

Bourdin, B., & Fayol, M. (2002). Even in adults, written production is still more costly than oral production. *International Journal of Psychology, 37*, 219-227.

Çelik, P., Storme, M., & Forthmann, B. (2016). A new perspective on the link between multiculturalism and creativity: The relationship between core value diversity and divergent thinking. *Learning and Individual Differences*, *52*, 188-196.

Chan, D., Schmitt, N., De Shon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology, 82*, 300-310.

Collins, E. C., Biernat, M., & Eidelman, S. (2009). Stereotypes in the communication and translation of person impressions. *Journal of Experimental Social Psychology, 45*, 368-374.

Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10*, 331-360.

Cottrell, J. M., Newman, D. A., & Roisman, G. I. (2015). Explaining the black–white gap in cognitive test scores: Toward a theory of adverse impact. *Journal of Applied Psychology, 100*, 1713-1736.

Crandall, C. S., Eshleman, A., & O'Brien, L. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. *Journal of Personality and Social Psychology, 82*, 359-378.

Cropley, A. (2006). In praise of convergent thinking. *Creativity Research Journal, 18*, 391-404.

Cucina, J. M., Chihwei, S., 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.

Dahlke, J. A., & Sackett, P. R. (2017). The relationship between cognitive-ability saturation and subgroup differences across predictors of job performance. *Journal of Applied Psychology, 102*, 1403-1420.

DeGroot, T., & Motowidlo, S. J. (1999). Why visual and vocal interview cues can affect interviewers' judgments and predict job performance. *Journal of Applied Psychology, 84*, 986–993.

De Soete, B., Lievens, F., Oostrom, J. K., & 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.

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. *Journal of Personality and Social Psychology, 82*, 835-848.

Edwards, B. D., & Arthur, W. (2007). An examination of factors contributing to a reduction in subgroup differences on a constructed-response paper-and-pencil test of scholastic achievement. *Journal of Applied Psychology, 92*, 794-801.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice (1978). *Uniform guidelines on employee selection procedures*: 29 C.F.R. 1607.

Fetzer, M., & Tuzinksi, K. (2014). *Simulations for personnel selection*. New York, NY: Springer.

Frazer, R. A., & Wiersma, U. J. (2001). Prejudice versus discrimination in the employment interview: We may hire equally, but our memories harbour prejudice. *Human Relations, 54*, 173-191.

Funke, U., & Schuler, H. (1998). Validity of stimulus and response components in a video test of social competence. *International Journal of Selection and Assessment, 6*, 115-123.

Goclowska, M. A., & Crisp, R. J. (2014). How dual-identity processes foster creativity. *Review of General Psychology, 18*, 216–236.

Goffin, R. D., Gellatly, I. R., Paunonen, S. V., Jackson, D. N., & Meyer, J. P. (1996). Criterion validation of two approaches to performance appraisal: The behavioral observation scale and the relative percentile method. *Journal of Business and Psychology, 11*, 23-33.

Gudykunst, W. B., Matsumoto, Y., Ting-Toomey, S., Nishida, T., Kim, K., & Heyman, S. (1996). The influence of cultural individualism-collectivism, self construals, and individual values on communication styles across cultures. *Human Communication Research, 22*, 510-543.

Guilford, J. P. (1950). Creativity. *American Psychologist, 5*, 444-445.

Hakel, M. D. (1998). *Beyond Multiple Choice: Evaluating Alternatives to Traditional Testing for Selection*. Mahwah, NJ: Lawrence Erlbaum Associates.

Harrison, D. A., Harrison, T., & Shaffer, M. A. (2018). Strangers in strained lands: Learning from workplace experiences of immigrant employees. *Journal of Management.* Advanced on line publication.

Hayes. A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach.* New York. NY: Guilford press.

Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? *American Psychologist, 47*, 1083-1101.

Hiemstra, A. M. F., Derous, E., Serlie, A. W., & Born, M. Ph. (2012). Fairness perceptions of video resumes among ethnically diverse applicants. *International Journal of Selection and Assessment*, *20*, 423-433.

Holder, K., & Kessels, U. (2017). Gender and ethnic stereotypes in student teachers' judgments: a new look from a shifting standards perspective. *Social Psychology of Education*, *20*, 471-490.

Hosoda, M., & Stone-Romero, E. F. (2010). The effects of foreign accents on employment-related decisions. *Journal of Managerial Psychology, 25,* 113-132.

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing, 26*, 219-244.

International Organization for Migration (2018). *World migration report.* Retrieved on March 4, 2018 from https://www.iom.int/wmr/world-migration-report-2018

International Task Force on Assessment Center Guidelines. (2009). Guidelines and ethical considerations for assessment center operations. *International Journal of Selection and Assessment, 17*, 243-253.

Iscoe, I., & Pierce-Jones, J. (1964). Divergent thinking, age, and intelligence in White and Negro children. *Child Development, 35*, 785-797.

Kaiser, C. R., & Pratt-Hyatt, J. S. (2009). Distributing prejudice unequally: Do Whites direct their prejudice toward strongly identified minorities? *Journal of Personality and Social Psychology, 96*, 432-445.

Kenny, D.A. (2015). *Measuring model fit.* Retrieved on March 19, 2018 from http://davidakenny.net/cm/fit.htm#null.

King, E. B., & Ahmad, A. S. (2010). An experimental field study of interpersonal discrimination toward Muslim job applicants. *Personnel Psychology, 63*, 881-906.

Klein, C., DeRouin, R. E., & Salas, E. (2006). Uncovering workplace interpersonal skills: A review, framework, and research agenda. In G. P. Hodgkinson & J. K. Ford (Eds.), *International review of industrial and organizational psychology* (Vol. 21, pp. 80-126). New York, NY: Wiley & Sons, Ltd.

Koch, A. J., D'Mello, S. D., & Sackett, P. R. (2015). A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Journal of Applied Psychology, 100,* 128-161.

Krumm, S., Lievens, F., Hüffmeier, J., Lipnevich, A.A., Bendels, H., & Hertel, G. (2015). How "situational" is judgment in situational judgment tests? *Journal of Applied Psychology, 100,* 399-416.

Landy, F. J. (2007). The validation of personnel decisions in the twenty-first century: Back to the future. In S. M. McPhail (Ed.), *Alternate validation strategies: Developing and leveraging existing validity evidence* (pp. 409-426). San Francisco, CA: Jossey-Bass.

Lenhard, W., Baier, H., Hoffmann, J., & Schneider, W. (2007). Automatic scoring of constructed-response items with latent semantic analysis. *Diagnostica, 53*, 155-165.

Leung, A. K. Y., Maddux, W. W., Galinsky, A. D., & Chiu, C. Y. (2008). Multicultural experience enhances creativity: The when and how. *American Psychologist*, *63*, 169-181.

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.

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.

Maddux, W. W., & Galinsky, A. D. (2009). Cultural borders and mental barriers: the relationship between living abroad and creativity. *Journal of Personality and Social Psychology*, *96*, 1047-1061.

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.

Minnaert, A. (1996*). Academic performance, cognition, metacognition and motivation. Assessing freshmen characteristics on task: A validation and replication study in higher education*. Unpublished doctoral dissertation, University of Louvain, Belgium.

Oostrom, J. K., & Born, M. Ph. (2014). Using cognitive pretesting to explore causes for ethnic differences on role-plays. *International Journal of Intercultural Relations, 41*, 138-149.

Oostrom, J. K., & De Soete, B. (2016). Ethnic differences in perceptions of cognitive ability tests: The explanatory role of self-serving attributions. *International Journal of Selection and Assessment*, *24*, 14-23.

Oostrom, J. K., Born, M. Ph., 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.

Oostrom, J. K., Born, M. Ph., 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.

Outtz, J., & Newman, D. A. (2009). A theory of adverse impact. In J. Outtz (Ed), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 53-94). New York, NY: Routledge.

Outtz, J., Goldstein, H., & Ferreter, J. (2006, April). *Testing divergent and convergent thinking: Test response format and adverse impact.* Paper presented at the 20th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology, 75,* 811-832.

Purkiss, S. L. S., Perrewé, P. L., Gillespie, T. L., Mayes, B. T., & Ferris, G. R. (2006). Implicit sources of bias in employment interview judgments and decisions. *Organizational Behavior and Human Decision Processes, 101*, 152-167.

Rindermann, H., Becker, D., & Coyle, T. R. (2016). Survey of expert opinion on intelligence: Causes of international differences in cognitive ability tests. *Frontiers in Psychology, 7,* 399.

Robie, C., Christiansen, N. D., Hausdorf, P. A., Murphy, S. A., Fisher, P. A., Risavy, S. D., & Keeping, L. M. (2017). International comparison of group differences in general mental ability for immigrants versus non-immigrants. *International Journal of Selection and Assessment, 25*, 347-359.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40*, 163-184.

Ryan, A. M. (2001). Explaining the Black-White test score gap: The role of test perceptions. *Human Performance, 14*, 45-75.

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: Lawrence Erlbaum Associates.

Schmid Mast, M., Gatica-Perez, D., Frauendorfer, D., Nguyen, L., & Choudhury, T. (2015). Social sensing for psychology: Automated interpersonal behavior assessment. *Current Directions in Psychological Science, 24*, 154-160.

Schmit, M. J., & Ryan, A. M. (1992). Test-taking dispositions: A missing link. *Journal of Applied Psychology, 77*, 629-637.

Simonton, D. K. (1999). *Origins of genius: Darwinian perspectives on creativity.* New York, NY: Oxford University Press.

Smith, K. C. & McDaniel, M. A. (1998, April). *Criterion and construct validity evidence for a situational judgment measure.* Paper presented at the 13th Annual Convention of the Society for Industrial and Organizational Psychology, Dallas. TX.

Te Nijenhuis, J., Willigers, D., Dragt, J., & Van der Flier, H. (2016). The effects of language bias and cultural bias estimated using the method of correlated vectors on a large database of IQ comparisons between native Dutch and ethnic minority immigrants from non-Western countries. *Intelligence*, *54*, 117-135.

Van Tubergen, F. (2010). Determinants of second language proficiency among refugees in the Netherlands. *Social Forces, 89*, 515-534.

VESOC (2003). *Ethnic minority criterion*. Retrieved on March 9, 2018 from http://www.serv.be/uitgaven/1330.pdf

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*, 557-574.

Weekley, J. A., & Jones, C. (1997). Video-based situational testing. *Personnel Psychology, 50*, 25-49.

Appendix A

The scoring of the constructed responses of the SJT of Study 1 was done in four stages. The first stage consisted of developing response clusters. After collapsing the original MC options as well as a sample of participants' written constructed responses (about 50% of the sample) into one item response pool, two trained coders (two female industrial and organizational psychology graduates; blind to the conditions) clustered the responses per item irrespective of response format. Discrepancies between the two coders were resolved through discussion with one of the authors. Across the items, 284 clusters were developed during this stage, with an average of nine clusters per item. For example, for the item "having to complete a difficult task", one cluster contained responses about asking help from colleagues, another cluster combined answers that one would attempt to find it out oneself, whereas still another cluster dealt with responses to ask help from one's supervisor(s), etc.

In the second retranslation stage, responses were assigned to the appropriate cluster. For composite constructed responses (e.g., "I would first talk to that colleague myself, but if the bullying does not stop I ask to meet with my supervisor") the number of responses was calculated (e.g., in this case: 2) and each of the responses was assigned to the corresponding response cluster. As inter-rater agreement for two coders in a subsample of 30 participants was satisfactory (average rater ICC[2,2] = .92, single rater ICC[2,2] = .86), the remaining cluster assignments (for 4,154 responses) were completed with one coder per response.

In the third stage, four employment coaches from the organization rated each of the response clusters on their effectiveness (from *1 = highly ineffective* to *5 = highly effective*). Given that the original MC responses had been sorted in the response clusters (see above), all potential responses (written constructed and MC responses) were scored by these subject matter experts. They were blind to the origin of the response clusters. Prior to developing the scoring key, they received instructions on how to define effective work behavior, practice items, and the performance standards in the employment agency. Two-way random intra-class correlations showed good inter-rater agreement (ICC[2,4] = .90).

Finally, all 31 responses of each participant were matched with the corresponding rating for the cluster to which the response was assigned. For composite responses consisting of multiple responses, an average score of the response ratings was calculated.

Appendix B

Table B1

Tests of Measurement Invariance for One-Factor Model Underlying SJT Scores Across MC and Written Constructed Response Formats in Study 1.

	X^2	*df*	ΔX^2	Δ*df*	*SRMR*	*RMSEA 95% CI*
Equal number of factors	940.521**	754			.078	.046 [.035-.055]
Equal factor loadings	978.190**	782	37.669	28	.086	.046 [.036-.055]

Note. *SRMR* = Standardized Root Mean Square Residual; *RMSEA* = Root Mean Square Error of Approximation. Given that the RMSEA of the null model was below Kenny's (2015) threshold of .1581, we did not include the TLI and CFI in the table.
** $p < .01$.

Table B2

Tests of Measurement Invariance for One-Factor Model Underlying Multimedia Test Scores across Response Formats in Study 2.

	X^2	*df*	ΔX^2	Δ*df*	*SRMR*	*RMSEA 95% CI*
Equal number of factors	232.121**	60			.081	.112 [.097-.127]
Equal factor loadings	253.623**	74	21.502	14	.093	.103 [.089-.117]

Note. *SRMR* = Standardized Root Mean Square Residual; *RMSEA* = Root Mean Square Error of Approximation. Given that the RMSEA of the null model was below Kenny's (2015) threshold of .1581, we did not include the TLI and CFI in the table.
** $p < .01$.

Table 1
Framework of Response Formats.

Categories	Example instructions	Prior research	Potential Mechanisms
Closed-ended (MC, forced-choice)	Pick the right response option; Rank order the response options; Rate the response options.	Arthur et al. (2014): Rate response format had least cognitive load and lower ethnic differences on integrity SJT (vs. rank and pick the best formats); Meta-analysis of Rodriguez's (2003): Scores on closed-ended formats had higher reliabilities than written constructed ones. Only construct equivalence when item stem was kept constant (see also meta-analysis of In'nami & Koizumi, 2009).	- Higher cognitive load
Written-constructed	Type the most effective response in the text box; Type how you would respond in the text box.	Funke and Schuler (1998): Higher validity for written constructed response SJT (vs. MC). Ethnic differences not examined; Edwards and Arthur (2007): Written constructed responses had 39% smaller ethnic differences on a knowledge test and more positive applicant reactions (vs. MC, see also Arthur, Edwards, & Barrett, 2002). Similar validities across formats.	- Lower cognitive load - Higher divergent thinking saturation - Higher test-taking motivation
Audio-visual constructed	Answer to the webcam as if you were to respond to the person.	Oostrom, Born, Serlie, and Van der Molen (2010, 2011) and Cucina, Chihwei, Busciglio, Harris Thomas, and Thompson Peyton (2015): Webcam video-based SJT was valid predictor of performance; De Soete, Lievens, Oostrom, and Westerveld (2013): Ethnic difference of $d = .14$ for a webcam video-based SJT; Lievens, De Corte, and Westerveld (2015): Webcam video-based SJT had higher validity for predicting training performance than written constructed response one. Ethnic differences not examined.	- Lower cognitive load - Rater effects

Note. As noted by Lievens and Sackett (2017), videoconference and face-to-face formats represent a fourth category. This category is not presented in this framework because it is not the focus of this study.

Table 2

Correlations Among Study 1 Variables for the Multiple Choice Condition.

	1.	2.	3.	4.	5.	6.
1. Ethnicity						
2. Home language	.75**					
3. Job experience	-.26**	-.30**				
4. Cognitive ability	-.42**	-.34**	.02			
5. Divergent thinking	-.27**	-.33**	.13	.24**		
6. Test motivation	-.03	.04	.14	-.04	.18*	
7. SJT	-.43**	-.39**	.12	.47**	.25**	.18*

Note. Ethnicity (0 = native; 1 = immigrant); Home language (0 = Dutch; 1 = Other). All other variables are in *z*-score format. *N* ranges from 134-137.
* *p* < .05, ** *p* < .01.

Table 3

Correlations Among Study 1 Variables for the Constructed Response Condition.

	1.	2.	3.	4.	5.	6.
1. Ethnicity						
2. Home language	.67**					
3. Job experience	-.20*	-.16				
4. Cognitive ability	-.34**	-.30**	.05			
5. Divergent thinking	-.30**	-.19**	.13	.33**		
6. Test motivation	.31**	.26**	-.08	-.18*	-.08*	
7. SJT	-.13	-.15	-.09	.14	.20*	-.09

Note. Ethnicity (0 = native; 1 = immigrant); Home language (0 = Dutch; 1 = other). All other variables are in *z*-score format. *N* ranges from 126-134.
* *p* < .05, ** *p* < .01.

Table 4

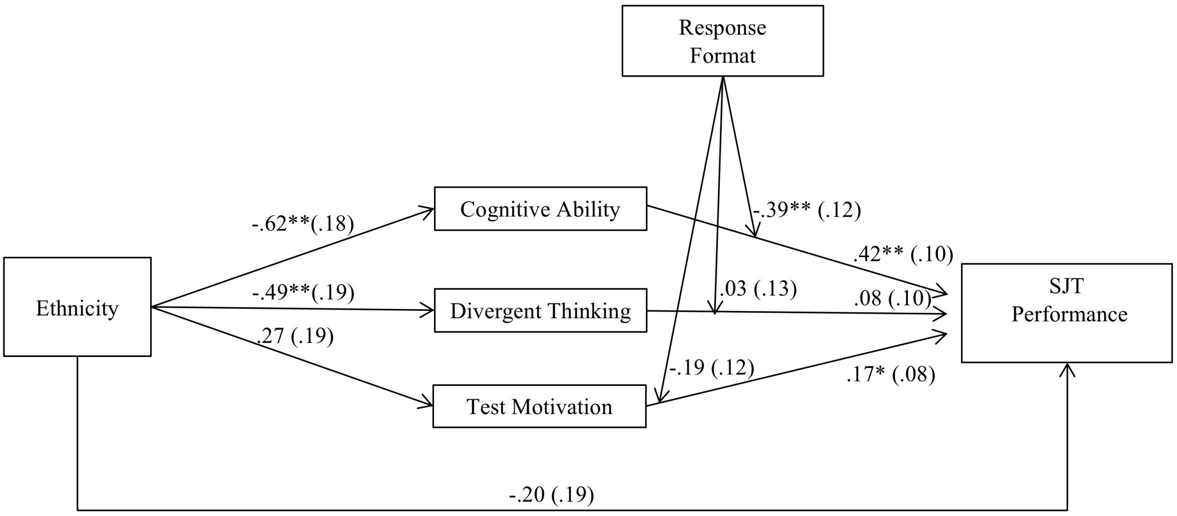
Correlations Among Study 2 Variables.

	1.	2.	3.	4.	5.	6.	7.
1. Ethnicity							
2. Home language	.71**						
3. Job experience	-.29**	-.27**					
4. Cognitive ability	-.40**	-.27**	.04				
5. SJT: Behavioral	-.03	-.03	.02	.16**			
6. SJT: Written	-.20**	-10	.13	.20**	.46**		
7. SJT: Transcribed	-.14*	-.14*	.09	.10	.84**	.55**	
8. Job performance	-.01	-.08	-.03	.04	.15*	.16*	.16*

Note. N ranges from 264 to 269, except for correlations with job performance, where *N* ranges from 192 to 193. Ethnicity (0 = native; 1 = immigrant); Home language (0 = Dutch; 1 = other). All other variables are in *z*-score format.

* *p* < .05, ** *p* < .01.

Figure 1
Results of Path Model of Study 1.



Conditional effect of cognitive ability on SJT:
.42** (.10) for MC
.03 (.08) for CR

Conditional effect of divergent thinking on SJT:
NA for MC
NA for CR

Conditional effect of test motivation on SJT:
NA for MC
NA for CR

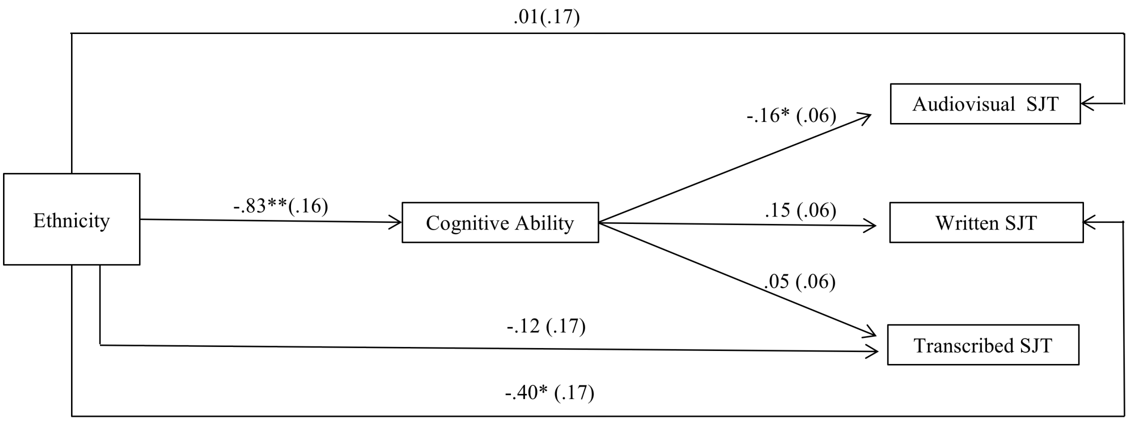
Indirect effect of ethnicity on SJT via cognitive ability:
-.26** (.09) for MC
-.02 (.06) for CR

Indirect effect of ethnicity on SJT via divergent thinking:
-.04 (.05) for MC
-.06 (.05) for CR

Indirect effect of ethnicity on SJT via test motivation:
.05 (.04) for MC
-.01 (.03) for CR

Note.
Index of moderated mediation = .24 (.06). Values in parentheses are bootstrapped SE's.
NA = As the Hayes (2018) SPSS PROCESS macro produces only conditional effects if there is a significant interaction, these values were not available. Response Format (MC = 0; Constructed response = 1). Ethnicity (0 = native; 1 = immigrant). For clarity, the controls (job experience and home language) are omitted from the figure. SJT performance, cognitive ability, divergent thinking, and test motivation are in *z*-score format. * $p < .05$, ** $p < .01$.

Figure 2
Results of Path Model of Study 2.



Indirect effects:		Total effects:	
Ethnicity – Audiovisual SJT:	-.13* (.06)	Ethnicity – Audiovisual SJT:	-.12 (.17)
Ethnicity – Written SJT:	-.12* (.06)	Ethnicity – Written SJT:	-.52**(.17)
Ethnicity – Transcribed SJT:	-.05 (.05)	Ethnicity – Transcribed SJT:	-.16 (.17)

Note.
Ethnicity (0 = native; 1 = immigrant). For clarity, the controls (job experience and home language) are omitted from the figure. SJT scores and cognitive ability are in *z*-score format.
* *p* < .05, ** *p* < .01.