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 versus 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 versus 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, situational judgment tests

According to the International Organization for Migration (IOM), 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, & Shaffer, 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, Su, Busciglio, Thomas, & Peyton, 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 (a) examining a range of response formats in terms of their effects on minority–majority differences and validity and (b) proposing and testing theoretical rationales underlying their effects (cognitive load, divergent thinking, test motivation, rater effects). Study 1 compares immigrants’ versus natives’ scores on an MC versus 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; Oitzl & Newman, 2009; see also Rindermann, Becker, & Coyle, 2016) developed a comprehensive model that distinguished between genetic, environmental (e.g., socioeconomic, 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, Crockley, 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; Golecwiska & 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 toward tests with an MC format: Minorities display lower test motivation and less favorable test perceptions, which then translates into lower test performance (Cham, Schmitt, DeShon, Clause, & Delbridge, 1997; Ryan, 2001; Schmit & Ryan, 1992). Although these results were obtained in the United State, 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 toward them (Hiemstra, Derous, Serlie, & Born, 2012). Similarly, in the United States, minorities showed more favorable attitudes toward 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

Method

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 years of job experience (vs. 9.6 years 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 European Union (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. Data were collected in accordance with the ethical guidelines for research involving human subjects by Ghent University, Belgium (1162012N).

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...
Table 2

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

Note. SJT = situational judgment test. 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.

The general cognitive ability. A time constraint of 22 min 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, and so forth. 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. 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. 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.

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.
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 and thus to lead to smaller SJT minority–majority differences for oral responses (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, reread, 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 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 documented the effect of shifting standards when evaluating minority versus majority members or males versus 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, Amrodio, 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, Civil Service Commission, Department of Labor, & Department of Justice, 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 rspected.

### Table 3

#### Correlations Among Study 1 Variables for the Constructed Response Condition

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Ethnicity</td>
<td>.67**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Home language</td>
<td>-.20*</td>
<td>-.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Job experience</td>
<td>-.34**</td>
<td>-.30**</td>
<td>.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Cognitive ability</td>
<td>-.30**</td>
<td>-.19**</td>
<td>.13**</td>
<td>.33**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Divergent thinking</td>
<td>-.26**</td>
<td>-.08</td>
<td>-.18</td>
<td>-.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Test motivation</td>
<td>.31**</td>
<td>.14</td>
<td>.20*</td>
<td>.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. SJT</td>
<td>-.13</td>
<td>-.15</td>
<td>-.09</td>
<td>.14</td>
<td>.20*</td>
<td>.09</td>
</tr>
</tbody>
</table>

*Note.* SJT = situational judgment test. 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.*
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 years of job experience (vs. 12.5 years 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, and so forth (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 1 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 interrater agreement (ICC[1,2] > .70), were they certified as assessors. The final assessor pool consisted of 24 industrial-organizational (I/O) psychology graduates (21 females; mean age = 23 years, 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 min. 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 ds 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 nonsignificant.

H5 stated that the relationship between SJT scores and the criterion is comparable for all three formats. As shown in Table 4, rs 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 (rs 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, ds 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 (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 min 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 versus 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, 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 benefits 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.

References


CONSTRUCTED RESPONSE FORMATS

Appendix A

Scoring of SJT

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), and so forth.

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 interrater 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 intraclass correlations showed good interrater 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.

(Appendices continue)
Appendix B

Summary of Measurement Invariance Tests

Table B1

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

<table>
<thead>
<tr>
<th>Tests</th>
<th>X^2</th>
<th>df</th>
<th>ΔX^2</th>
<th>Δdf</th>
<th>SRMR</th>
<th>RMSEA 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equal number of factors</td>
<td>940.521**</td>
<td>754</td>
<td></td>
<td></td>
<td>.078</td>
<td>.046 [.035—.055]</td>
</tr>
<tr>
<td>Equal factor loadings</td>
<td>978.190**</td>
<td>782</td>
<td>37.669</td>
<td>28</td>
<td>.086</td>
<td>.046 [.036—.055]</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Tests</th>
<th>X^2</th>
<th>df</th>
<th>ΔX^2</th>
<th>Δdf</th>
<th>SRMR</th>
<th>RMSEA 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equal number of factors</td>
<td>232.121**</td>
<td>60</td>
<td></td>
<td></td>
<td>.081</td>
<td>.112 [.097—.127]</td>
</tr>
</tbody>
</table>

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.
AUTHOR PLEASE ANSWER ALL QUERIES

AQau—Please confirm the given-names and surnames are identified properly by the colors. 

<table>
<thead>
<tr>
<th>Color</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>Given-Name</td>
</tr>
<tr>
<td>Blue</td>
<td>Surname</td>
</tr>
</tbody>
</table>

The colors are for proofing purposes only. The colors will not appear online or in print.

AQ1—Author: Please confirm that running header is acceptable or provide alternate header of 50 characters max. including spaces.

AQ2—Author: In the title, “Minority–Majority” is separated with an en-dash rather than a hyphen to indicate relationship between words. Please check that this punctuation edit to the title is acceptable.

AQ3—Author: Please be sure to provide the name of the department(s) with which you and your coauthors are affiliated at your respective institutes if you have not already done so. If you are affiliated with a governmental department, business, hospital, clinic, VA center, or other nonuniversity-based institute, please provide the city and U.S. state (or the city, province, and country) in which the institute is based. Departments should be listed in the author footnote only, not the byline. If you or your coauthors have changed affiliations since the article was written, please include a separate note indicating the new department/affiliation: [author’s name] is now at [affiliation].

AQ4—Author: Please check if location for non-academic institutional affiliation Cubiks Belgium is correct.

AQ5—Author: Please limit the number of Keywords or phrases to a maximum of 5.

AQ6—Author: Citation changed to (Hiemstra, Derous, Serlie, & Born, 2012) to match reference. Please confirm this is correct.

AQ7—Author: Please check if correct as edited: “...24 industrial-organizational (I/O) psychology graduates...”

AQ8—Author: Please provide an appropriate title for the appendix.

AQ9—Author: Please provide an appropriate title for Appendix B that describes both Tables.

AQ10—Author: Please provide a header for Tables B1 and B2 in Appendix B.

AQ11—Author: Please check if location for non-academic institutional affiliation Cubiks Belgium is
AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

correct