Journal of Business and Psychology (2021) 36:693–711
https://doi.org/10.1007/s10869-020-09697-1

ORIGINAL PAPER

Performance on Video-Based Situational Judgment Test Items: Simulated Interracial Interactions

Juliya Golubovich[1] · Ann Marie Ryan[1]

Published online: 5 June 2020

Abstract

Individuals are known to categorize others into social groups based on cues like race and gender and to experience relative discomfort when interacting with "outgroup" members. Two experimental studies were used to examine whether actor demographic cues in situational judgment assessment items completed by test takers in a simulated employee selection context may lead to differences in their performance and reactions to the hiring organization. In both studies, test takers assumed the perspective of actors shown in video-based scenarios and indicated how they would respond to interaction partners (IPs) to whom they were racially similar or dissimilar. In Study 1, a given test taker responded to IPs of a constant gender; in Study 2, IPs' gender varied across scenarios within each condition. In Study 1, Black test takers spent more time and scored better on two of the four scenarios when responding to racially similar IPs. These effects were not found in Study 2, but demographic cues showed new interactive effects on performance and reactions. We discuss the implications of different findings across the two studies.

Keywords Situational judgment tests · Interracial interactions · Social categorization · Organizational recruitment · Applicant reactions

Selection tool developers have long been concerned with making sure the content of assessments does not lead to construct irrelevant variance affecting scores (Guion 1998). As the capacity for richer media use in assessment content delivery has increased (i.e., video items, animation), one concern is whether these enhancements affect performance and engagement (e.g., Chan and Schmitt 1997) or add noise or systematic error (Hawkes et al. 2018). An international survey of HR professionals found that the majority of respondents' organizations already used or intended to start using video/multimedia in assessments (Ryan et al. 2015). Consequently, there is a need to consider how enhanced fidelity may influence job candidates' performance and reactions.

Relational demography research, which involves the comparison of demographic profiles of individuals who interact with one another (Tsui et al. 1992), provides an established framework for investigating one particular area where rich media differ from traditional paper and pencil assessments and that is in the provision of social category information within test items. Relational demography research proposes that individuals tend to be attracted to similar others and that this has implications for individuals' interaction quality and personal outcomes (e.g., Graves and Elsass 2005; Goldberg et al. 2008). Given that the depictions of interpersonal interactions in media-rich assessment provide demographic information on interaction partners that is not prototypically part of less-rich formats, relational demography research suggests this may add a systemic source of variance in performance.

Furthermore, research on identity contingency cues, which are indicators of whether an environment might be psychologically safe for an individual (Murphy et al. 2007), suggest that encountering demographic information in the context of an assessment may influence perceptions of fit and belonging (Murphy and Taylor 2012). The provision of demographic

This research was conducted while the first author was a graduate student at the Michigan State University.

✉ Juliya Golubovich
jgolubovich@gmail.com

Ann Marie Ryan
ryanan@msu.edu

[1] Michigan State University, 316 Physics Rd., East Lansing, MI 48824, USA

information in media-rich assessments may serve as an identity contingency cue for job candidates.

The current research draws from these findings on relational demography and identity contingency cues to examine the influence of demographic cues in video situational judgment assessment (SJT) items. SJTs are a measurement method that involves asking test takers to consider and respond to hypothetical situations (Christian et al. 2010). The hypothetical scenario and its set of response options are typically collectively referred to as an "SJT item." SJT scenarios are typically sampled from challenging work situations that are often interpersonal in nature (Campion et al. 2014). Test takers may be asked to pick a best/worst response, their most likely/least likely response, or to rate all the response options on their effectiveness or likelihood (Campion et al. 2014). Video-based SJTs show test takers the scenarios (versus presenting them in written, paper-and-pencil form), as this can improve assessment fidelity, validity, fairness, and reactions (Chan and Schmitt 1997; Weekley and Jones 1997). However, video-based SJTs also tend to have lower internal consistency,[1] likely due to the extra information (item-specific variance) present in the video that is absent from written scenarios (Campion et al. 2014). The demographics of interaction partners in video-based scenarios may be one contributor to this variance because video SJTs frequently feature interpersonal interactions. Workplace situations involving interactions are common for jobs involving teamwork or customer service and they lend themselves well to video—*showing* interactions provides non-verbal information, contextual information, and greater realism (for examples of such SJTs, see de Meijer et al. 2010, Lievens 2013, and MacCann et al. 2016).

Across two studies, we examine the effects of test takers' racial similarity to actors simulating coworkers in interpersonally challenging scenarios that are prototypical of SJT items. We also examine the role of actors' gender in these effects. We provide a contribution to the literature and practice on selection assessment and also to theory and research on identity contingency cues in evaluation contexts (Murphy and Taylor 2012; Purdie-Vaughns et al. 2008) in two ways. First, while assessment developers often attempt to signal a positive climate for diversity via depicting diverse individuals in assessment materials (e.g., using diverse pronouns and names in items; depicting diverse individuals in pictures and videos), we examined whether the race and gender of individuals in assessment materials influence test taker performance. As we will discuss shortly, the literature on interracial interactions clearly suggests that test takers might respond differently to imagined interactions with video/virtual ingroup versus outgroup members, particularly in the types of challenging workplace scenarios that often serve as content for selection assessments. Given the literature on intersectionality (Crenshaw 1989), interaction partners' gender may moderate such effects. Second, while research on identity contingency cues has explored how the representation of diverse individuals in recruitment materials may signal belongingness and identity safety to minority job applicants (e.g., Avery et al. 2004; Walker et al. 2012), this stream of research has not considered portrayal in assessment content as a potential identity contingency cue. We examine test taker reactions to diversity within assessment content.

After providing a review of relevant research, we propose how the depictions of demographic cues in video-based SJTs might influence responses and reactions to SJT scenarios. Subsequently, we present two studies designed to assess these effects. Both studies manipulate racial and gender cues in SJT scenarios, but Study 2 presents each test taker with a greater level of gender diversity than Study 1.

Social Categorization

According to social categorization theory (Tajfel and Turner 1979) and empirical findings (Ito and Urland 2003; Montepare and Opeyo 2002), individuals characterize others on the basis of demographic characteristics like race and gender quickly and automatically. Such social categorizations are believed to serve as cognitive tools that help individuals organize and make sense of their social environment in order to know how to behave (Tajfel and Turner 1979). Categorizing others into social groups allows individuals to demonstrate a preference for ingroups ("us") relative to outgroups ("them") and reinforces their own social identities in the process (Tajfel and Turner 1979). Relatedly, the similarity-attraction paradigm (Lau et al. 2008), upon which relational demography research is based, proposes that individuals are more attracted to similar others.

Social categorization processes are particularly informative for understanding initial interactions with "unknown" others, about whom individuals have little individuating information (Rothbart and John 1985). In this context, individuals use cues like race and gender to make assumptions about others' attitudes, values, and behaviors (West and Dovidio 2013). This includes the automatic activation of stereotypes held about the demographic groups to which these individuals belong (Devine 1989). These types of interactions are of interest in the current

[1] Campion et al. (2014) reviewed SJT research and correlated SJT attributes with their coefficient alpha values. Alpha values were lower when video versus written SJT formats were used ($r = -.28$, $p < .05$).

study, where scenarios present interactions with actors with whom applicants are not familiar.[2]

Interracial Interactions and Test Performance

The Role of Race For historically privileged majority group members, the prospect of interacting with members of a different racial group is likely to trigger concerns about the possibility of appearing prejudiced (Shelton et al. 2010; Trawalter et al. 2009). The reasons have to do with historic inequalities between racial minority and majority groups and consequent societal efforts to promote the treatment of all individuals in a "colorblind" manner (Ryan et al. 2007). Individuals who are concerned about appearing prejudiced during interactions with outgroup members may try to monitor their thoughts, feelings, and behaviors for signs of prejudice (Dovidio and Gaertner 2004; Richeson and Shelton 2003). Although racial minorities may also worry about appearing prejudiced during interracial interactions, they are particularly likely to be apprehensive about finding themselves on the receiving end of prejudice (Doerr et al. 2011; Mendoza-Denton et al. 2002; Pinel 1999; Shelton and Richeson 2006). Racial minorities worried about encountering bias may monitor their majority interaction partners' behavior for signs of negative attitudes (Major et al. 2002; Vorauer 2006). Relatedly, researchers have demonstrated minorities' sensitivity to cues indicative of racial bias (Richeson and Shelton 2004; Wout et al. 2014).

Concerns about appearing prejudiced or being subjected to prejudice are associated with anxiety for those thinking about, expecting, or briefly engaging in interracial interaction (e.g., Mendoza-Denton et al. 2002; Plant 2004; Plant and Butz 2006; Shelton 2003). Interracial interactions may be more cognitively taxing because of the need to exert cognitive resources to self-monitor, inhibit certain behaviors, and monitor an interaction partner's behavior (Apfelbaum et al. 2008; Plant and Butz 2006). Anxious thoughts and other cognitive distractions can cause task performance to suffer because they consume limited attentional resources (Moran 2016; Randall et al. 2014). Even a short interracial interaction may temporarily deplete attentional resources for both minority and majority group members (Richeson and Trawalter 2005; Richeson and Shelton 2003; Richeson et al. 2005).

Given that even anticipated interracial interactions have been shown to produce anxiety, we expect that imagined interactions with actors of a different race shown in a video-based SJT can trigger the same types of concerns about appearing prejudiced or being subjected to prejudice. Although test takers may not necessarily expect to be evaluated or stigmatized by their interaction partners in imagined interactions, they know their responses to these interactions will be judged. Thus, while in a simulated engagement with actors of a different race, test takers' prejudice-related concerns may not only interfere with cognitive processing of the situation and response options but also cause individuals to change the lens through which they evaluate the potential responses (e.g., Will responding this way make me appear prejudiced? Would I be stigmatized for responding this way?). Thus, individuals' responses to video-based SJT scenarios may be influenced by their similarity to the actor with whom they imagine interacting (i.e., similarity to the actor may affect individuals' ability to respond "correctly"). We predict that:

> H1: Test taker-interaction partner racial similarity will be positively associated with performance on SJT items.

Although interracial interactions can be anxiety provoking and distracting for both White and minority group members, meta-analytic findings suggest that minorities may be able to better manage interracial interactions than majority group members (Toosi et al. 2012). As a function of numbers, minority group members tend to have more experience interacting with White (majority) group members than Whites have experience interacting with them (Vorauer 2006). Minorities may be able to engage in compensatory strategies that help them have relatively pleasant interactions with Whites whom they expect to be prejudiced (Shelton et al. 2005); these are the types of skills that come with experience (Stephan and Stephan 1985). We do not wish to imply that individuals always have the resources to cope with prejudice adaptively or that experiencing racism does not have serious consequences (for a review of these issues, see Clark et al. 1999). However, we anticipate that the effects of racial similarity may vary by respondent's race:

> H2: Test taker race will moderate the effects of test taker-interaction partner racial similarity on SJT item performance, such that the positive effect of test taker-interaction partner racial similarity will be stronger for White test takers relative to Black and Asian test takers.

Concerns about appearing prejudiced or being subjected to prejudice may be associated with a greater desire to avoid or

[2] Brief interactions with or observations of strangers differ from the more sustained interactions over time examined in the context of cross-group friendships (e.g., West and Dovidio 2013). Friendship development involves more complex and dynamic processes where interaction partners influence each other and evidence change in personal attitudes and behavior over time (West and Dovidio 2013). For example, as individuals get to know outgroup members, they can find out individuating information that may be inconsistent with their held stereotypes (Rothbart and John 1985). Thus, we bound our literature review and theoretical rationale to contexts of interracial interactions between individuals who are new interaction partners or who have less close relationships and not to discussions of long-standing, well-developed close relationships.

quickly terminate interracial interactions (Doerr et al. 2011; Plant 2004; Plant and Butz 2006; Wout et al. 2014). On the other hand, distracted individuals may spend more time on interracial than similar race interactions as they try to decide on appropriate ways to respond to the posed situations. Thus, we also examine the main effect of test taker-interaction partner racial similarity (RQ1) and its interactive effect with test taker race on time spent responding to the scenarios (RQ2).

The Role of Gender In the context of interpersonal situations, individuals' gender can influence how they are perceived and treated by others (e.g., Balliet et al. 2011; Carli 1989; Russell and Owens 1999; Shute and Charlton 2006). For example, negative attitudes toward outgroup members tend to be targeted toward the males of the group more so than the females (Toosi et al. 2012). Thus, researchers recommend simultaneously considering the meaning and consequences of multiple categories of identity ("intersectionality"; Crenshaw 1989/1993) and relational demography research frequently examines gender as a moderator (e.g., Kirchmeyer 1995; Tsui et al. 1992). Therefore, we also check for moderating effects of interaction partners' gender on test takers' SJT performance (RQ3) and scenario response time (RQ4).

Diversity and Fit Perceptions

In addition to being used as a selection tool, SJTs are also viewed as a way to give candidates a realistic job preview (e.g., Campion et al. 2014). Candidates may decide they are well suited to dealing with the types of situations characteristic of the job or choose to self-select out of the process instead (O'Connell et al. 2013). As such, SJTs may function as a recruitment tool that helps attract candidates with better fit for the job. Furthermore, characteristics of an SJT are expected to influence applicants' reactions (Bauer and Truxillo 2006).

Researchers in the area of targeted recruiting have devoted substantial attention to examining the determinants of minority applicants' reactions to recruitment tactics (Avery and McKay 2006). Minorities (Blacks, Hispanics) react positively to racial diversity in advertisements while Whites are not influenced by the level of diversity in ads (Avery et al. 2004; Perkins et al. 2000). Employees in company materials, like a video-based SJT used to assess candidates, may be viewed as proxies for actual employees, whom individuals are not able to observe, and used to judge potential fit with the organization from a demographic standpoint (Perkins et al. 2000). It may be the case that White applicants, by virtue of their status as majority group members, take it for granted that their demographic group will be represented and treated well in organizations and therefore do not pay close attention to diversity cues in recruitment materials; minority applicants may need to be more attuned to this type of information to predict how well they will be represented and treated (Avery 2003; Perkins et al. 2000). Social psychology researchers have established that lower representation of one's minority group triggers a lowered sense of belonging and less desire to engage (Murphy et al. 2007). Furthermore, there is some evidence that minorities may judge diversity based on minority group representation, as opposed to simply own group representation, reacting positively as long as some minority group (e.g., Black, Hispanic) is well represented (Avery et al. 2004). Research on cue transfer suggests an explanatory mechanism—minority groups use diversity cues to judge organizations' level of egalitarian values and treatment of different groups as being equal, so cues that suggest safety for one group can also imply safety for another (e.g., Chaney and Sanchez 2018; Chaney et al. 2016). The highlighted research would suggest that minorities may react positively to diversity in a video-based SJT, whereas Whites may be indifferent. Thus, we predict that:

> H3: Test taker race will moderate the effects of SJT scenario diversity on test takers' fit perceptions such that the effects of SJT scenario diversity will be positive for Black and Asian test takers, but not affect White test takers' fit perceptions.

We conducted two studies where we manipulated the race of test takers' interaction partners across conditions (i.e., at the between-subject level), but kept it constant across video-based SJT scenarios (i.e., at the within-subject level). We varied the gender of the interaction partner across conditions in Study 1 and across scenarios in Study 2 to explore the relative salience of race and gender cues more thoroughly.

Study 1 Method

Sample

Undergraduate students at a large Midwestern University participated in the study in exchange for partial credit in their psychology courses (final $N = 335$). Females comprised 51.6% of the sample; 78.8% were White, 12.2% were Asian, and 9.0% were African American/Black.[3] Their average age was 19.88 (SD = 2.64). Finally, they were 26.3% freshmen, 31.3% sophomores, 22.4% juniors, 19.7% seniors, and 0.3% other/non-degree.

[3] As hypothesis testing was based on White, Asian, and Black actors, additional participants who were not from those demographic groups ($N = 21$) were excluded from analyses.

Measures and Design

SJT Items We selected four scenarios from a longer, validated test formerly used by a major manufacturer as part of its selection process for plant technicians to ensure that our items were representative of the types of SJT items used in employee selection for entry-level jobs. Two situations involved one coworker asking another for help or a favor and the other two involved a conflict between coworkers. Employees' behavior in helping and conflict situations is related to outcomes organizations value (e.g., Podsakoff et al. 2000; Tjosvold 1998), and these situations are commonly found in SJTs (e.g., Chan and Schmitt 2002; Olson-Buchanan et al. 1998; Oswald et al. 2004). Working with the scenario scripts, we professionally videotaped the four scenarios with different actors to create our SJT stimuli for six different experimental conditions.

For each situation, test takers watched two actors (playing coworkers) act out a situation which then required a response from one of the actors (referred to in this paper as "the assumed role actor" because test takers were then asked to assume that person's role when selecting how they would respond to the IP). The length of videos ranged from 13 to 50 s per scenario. Across and within conditions, the assumed role actor was always a White female.[4] Within a given experimental condition, different assumed role actors appeared in the four scenarios; across conditions, the assumed role actor in a given scenario was always the same White female (see Table 1 for an overview of the experimental design). The race (Asian, Black, or White) and gender (male, female) of the assumed role actor's coworker ("interaction partner"), whom respondents imagined interacting with, varied across conditions. Thus, considering a given SJT scenario, regardless of experimental condition, it was always "Ann" (same White female) shown interacting with an IP, an actor who varied in race/gender depending on the experimental condition the participant was in. Within a condition, interaction partners ("IPs") across the four scenarios were of the same race and gender. Every IP (e.g., Asian male) appeared in one scenario in one particular condition. Actors were 18–30 years old.

After a given scenario video ended, test takers were presented with a set of five or six text-based response options. They indicated which option they would *most likely* and *least likely* do if they were in the position of the assumed role actor (e.g., "Ann"). A *most likely* (*least likely*) response was assigned a score of 1 if it matched what the company considered to be a desirable (undesirable) response to the situation, a score of −1 if it matched a response considered undesirable (desirable), and a score of 0 if it was a match for neither a desirable nor an undesirable response. *Most likely* and *least likely* scores were summed. Thus, an item score could range from −2 (if someone selected an undesirable answer as their most likely response and a desirable answer as their least likely response) to +2 (if someone selected a desirable answer as their most likely response and an undesirable answer as their least likely response).

The manufacturer's scoring key for these items was based on response frequencies obtained from a sample of incumbent technicians and was reviewed by personnel at the company to ensure consistency with performance expectations for plant technicians. The manufacturer's full-length video SJT was validated against job performance data of approximately 400 plant technicians ($r = .26$ after correction for restriction of range in assessment scores and reliability of performance ratings).

Fit Perceived fit ($\alpha = .91$) was assessed with 10 items adapted from Perkins et al.'s (2000) attraction (e.g., "I would speak to a company representative about the possibility of employment") and compatibility ("I would feel at home working for an organization like this") scales. These scales were combined because they were correlated at above .70. All answers were provided on a five-point scale ranging from (1) *strongly disagree* to (5) *strongly agree*.[5]

Procedure

Up to eight students at a time participated in a supervised data collection at a computer lab. After providing informed consent, they were asked to imagine themselves as job applicants taking an interpersonal test as an initial step in the process of applying for a technician position at "ABC Corporation." The company was described as offering a high salary and good career growth, being located in a favorable location, and being rated a top place to work. A short description of the technician job was also included. Participants then completed the SJT scenarios

[4] Research on avatar gender has shown that as many as 79% of those in multiplayer online games have used avatars of an opposite gender and 30% do so on a regular basis (Hussain and Griffiths 2008; Yee & Ducheneaut, 2011, as cited by Martey et al. 2014). Findings fairly consistently show that when assuming an avatar of a different gender people do not tend to behave differently (i.e., movement within a game), but when men take on a female role, they tend to adopt stereotypic language patterns in chat functions (i.e., use more emotional language, use more exclamation points; Yee and Bailenson 2007; Martey et al. 2014). As our response options were not gendered in language patterns and individuals were responding to a multiple choice item, this research would suggest that gender assumed should not have an effect. Men and women in the current research generally did not perform differently on the SJT scenarios.

[5] Two measures of reactions to the SJT scenarios—perceived opportunity to perform and job relatedness—were also included in the current studies, but there were no differences in these measures across conditions in either Study 1 or Study 2. In the interests of brevity, these measures are not discussed. Ratings of interaction partner attractiveness were also collected. As these ratings were generally uncorrelated with the outcomes of interest and controlling for these ratings in testing the hypotheses and research questions in studies 1 and 2 did not substantively change our results, these measures are likewise not discussed.

Table 1 Study 1 design

	Condition 1		Condition 2		Condition 3		Condition 4		Condition 5		Condition 6	
SJT item	ARA	IP	ARA	IP	ARA	IP	ARA	IP	ARA	IP	ARA	IP
Conflict 1	ARA1	AF1	ARA1	AM1	ARA1	BF1	ARA1	BM1	ARA1	WF1	ARA1	WM1
Helping 1	ARA2	AF2	ARA2	AM2	ARA2	BF2	ARA2	BM2	ARA2	WF2	ARA2	WM2
Conflict 2	ARA3	AF3	ARA3	AM3	ARA3	BF3	ARA3	BM3	ARA3	WF3	ARA3	WM3
Helping 2	ARA4	AF4	ARA4	AM4	ARA4	BF4	ARA4	BM4	ARA4	WF4	ARA4	WM4

Note: *ARA*, assumed role actor, a White female. Assumed role actors indicated with the same number across columns are the same individuals. *IP*, interaction partner; *AF(M)*, Asian female (male); *BF(M)*, Black female (male); *WF(M)*, White female (male). A given actor is identified with the same acronym and number in study 1 and study 2 (see Table 5). Scenarios were presented in random order

(with random assignment to one of six conditions), followed by the fit measure. The four SJT scenarios were shown in random order. Time spent viewing the SJT scenario videos (prior to responding) was constant across participants because the next screen did not appear until the full video finished playing. Time spent responding to the four scenarios varied; an overall scenario response time was calculated by summing the individual scenario response times.

Study 1 Results

Data Cleaning

Removing inattentive respondents is recommended to enhance the validity of research findings (Maniaci and Rogge 2014). Common recommendations are to remove individuals who failed attention checks and/or provided data with multivariate outliers (e.g., Meade and Craig 2012). We excluded ten participants who failed the attention check for one or more scenarios by choosing the same option as their most and least likely response to the scenario and two additional participants who provided data with multivariate outliers.[6] Outliers were identified via Mahalanobis distance (Mahalanobis 1936), which considers response patterns across a series of items and flags individuals furthest from the average response vector. A probability level of $p < .001$ was used. The final sample size was 335.

Descriptive Statistics

Intercorrelations, means, standard deviations, and reliabilities are provided in Table 2. Male respondents spent more time on the scenarios ($r = .16, p < .01$) and perceived better fit with the organization ($r = .15, p < .01$). Those who perceived better fit with the organization performed better ($r = .27, p < .001$) and spent more time on the scenarios ($r = .12, p < .05$).

Hypothesis 1

To test for a positive effect of test taker-IP racial similarity on test taker performance, we conducted a general linear model (GLM) with test taker-IP racial similarity as the independent variable and scores on the four SJT items as the dependent variables. Test takers did not perform better on the scenarios when interacting with racially similar IPs, Wilk's Lambda = .98, $F(4{,}327) = 1.85$, ns. Thus, hypothesis 1 was not supported.

Research Question 1

To examine the effects of test taker-IP racial similarity on response time, we conducted a GLM with racial similarity as the independent variable and response times to the four SJT items as the dependent variables. The amount of time test takers spent responding to the scenarios did not differ as a function of test taker-IP racial similarity, Wilk's Lambda = 1.00, $F(4{,}327) = 0.24$, ns.

Hypothesis 2

To examine whether test taker race moderates the effects of test taker-IP racial similarity on performance, we conducted a GLM with racial similarity and test taker race (Asian, Black, or White) as independent variables, and scores on the four SJT items as the dependent variables. The interaction was statistically significant, Wilk's Lambda = .95, $F(8{,}646) = 2.12$, $p < .05$.

As shown in Table 3, Asian and White test takers did not perform better on the items when interacting with racially similar IPs. Black test takers performed better in helping

[6] The pattern of significant results did not change when rerunning analyses without excluding multivariate outliers.

Table 2 Descriptives and intercorrelations for Study 1 variables

	M	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Scenario diversity	0.69	0.46	-																
2. IP race	0.49	0.50	--[a]	-															
3. IP gender	0.49	0.50	0.02	− 0.04	-														
4. Race similarity	0.34	0.47	− 0.65	−0.04	− 0.06	-													
5. CS1 score	1.24	0.95	− 0.02	− 0.01	− 0.05	0.04	-												
6. CS2 score	0.56	0.84	0.07	− 0.02	− 0.08	− 0.01	0.05	-											
7. HS1 score	0.61	0.95	− 0.06	0.11	0.09	0.10	0.05	0.06	-										
8. HS2 score	0.65	0.59	0.01	0.03	− 0.01	0.11	0.06	0.04	0.03	-									
9. Total score	3.07	1.80	− 0.01	0.05	− 0.02	0.11	0.60	0.54	0.59	0.39	-								
10. CS1 time	1.00	0.37	0.06	0.12	− 0.05	− 0.03	0.00	0.02	0.06	0.02	0.05	-							
11. CS2 time	0.89	0.35	0.06	0.02	0.01	0.02	0.01	− 0.08	0.01	0.01	− 0.02	0.45	-						
12. HS1 time	0.94	0.37	0.06	0.06	− 0.04	− 0.02	0.03	0.04	− 0.05	0.01	0.01	0.39	0.58	-					
13. HS2 time	0.88	0.34	0.11	0.03	0.01	− 0.01	0.12	0.04	− 0.01	0.06	0.10	0.45	0.54	0.56	-				
14. Total time	3.71	1.12	0.09	0.08	− 0.03	− 0.01	0.05	0.01	0.00	0.03	0.04	0.73	0.81	0.80	0.80	-			
15. Fit	3.42	0.79	− 0.02	− 0.01	0.00	0.05	0.15	0.12	0.16	0.15	0.27	0.14	0.07	0.08	0.08	0.12	0.91		
16. Participant gender	0.48	0.50	0.00	0.00	− 0.01	− 0.04	0.04	0.08	− 0.07	− 0.04	0.01	0.06	0.19	0.15	0.09	0.16	0.15	-	
17. Participant race	0.79	0.41	− 0.02	− 0.01	− 0.01	− 0.10	0.02	0.01	0.01	0.11	0.06	− 0.07	− 0.19	− 0.19	− 0.12	− 0.18	0.04	− 0.04	-

Note: $N = 231–335$. *IP*, interaction partner; *CS*, conflict situation; *HS*, helping situation. For scenario diversity, 0 = White interaction partners and 1 = Asian or Black interaction partners. For IP race, 0 = Asian and 1 = Black. For IP gender, 0 = female and 1 = male. For race similarity, 0 = dissimilar to interaction partners and 1 = similar. For participant gender, 0 = female and 1 = male. For participant race, 0 = Asian or Black and 1 = White. Time variables are in minutes.

Correlations in italics are significant at the $p < .05$ level or better. Reliability for the fit measure is shown on the diagonal.

[a] Correlation cannot be computed because the two variables represent different coding of the race manipulation

Table 3 Performance on SJT items as function of test taker-IP racial similarity, study 1

Scenario	Test taker race	Racially similar IP *M* (SD)	Racially dissimilar IP *M* (SD)	*t* test	*d*
Helping 1	Asian	0.76 (1.03)	0.59 (1.10)	*t*(37) = − 0.50, ns	0.16
	Black	0.77 (1.01)	0.24 (1.09)	*t*(28) = − 1.37, ns	0.51
	White	0.72 (0.90)	0.56 (0.92)	*t*(261) = − 1.38, ns	0.18
Helping 2	Asian	0.71 (0.59)	0.36 (0.58)	*t*(37) = − 1.81, *p* = .078	0.59
	Black	0.92 (0.76)	0.18 (0.53)	*t*(28) = − 3.18, *p* < .01	1.14
	White	0.71 (0.51)	0.67 (0.60)	*t*(261) = − 0.51, ns	0.07
Conflict 1	Asian	0.94 (0.97)	1.00 (1.27)	*t*(37) = 0.56, ns	0.05
	Black	1.62 (0.77)	1.29 (0.69)	*t*(28) = − 1.21, ns	0.44
	White	1.31 (0.92)	1.22 (0.94)	*t*(261) = − 0.73, ns	0.10
Conflict 2	Asian	0.41 (1.00)	0.50 (0.96)	*t*(37) = 0.28, ns	0.09
	Black	1.08 (0.64)	0.41 (0.87)	*t*(28) = − 2.31, *p* < .05	0.76
	White	0.51 (0.83)	0.59 (0.82)	*t*(261) = 0.81, ns	0.11

situation 2 and conflict situation 2 when interacting with racially similar IPs.[7]

To understand the nature of the score differences, we cross tabulated the response variables with racial similarity. The helping situation involved an IP asking the assumed role actor to take on an inconvenience for a work-related reason, though the assumed role actor believes the IP is just trying to avoid the undesirable work (she points this out to the IP). Black test takers selected the keyed “best” response of saying no and suggesting another way for the IP to resolve the problem for their “most likely” choice and the keyed “worst” response of confronting the coworker regarding taking advantage for their “least likely” choice somewhat more frequently after watching a racially similar IP. Additionally, those who watched a racially dissimilar IP tended to select a neutrally keyed option of agreeing to the request with a quid pro quo as their “most likely” choice. That is, Black test takers tended to choose a highly agreeable response when the IP was racially dissimilar, and more of a direct or confrontation response when the IP was racially similar.

The conflict situation involved the IP pointing to the assumed role actor’s mistakes as a source of conflict. Black test takers selected the keyed “best” response of asking to meet with the IP and their supervisor to discuss differences for their “most likely” choice and the keyed “worst” response of avoiding the IP for their “least likely” choice somewhat more frequently after watching a racially similar IP. Once again, a more direct, confrontation response was chosen less when the IP was racially dissimilar.

No significant differences as a function of racial similarity were observed for conflict situation 1 or helping situation 1. Since the effects of racial similarity were more positive for Black relative to White and Asian test takers (for the two scenarios discussed), the nature of the interaction between racial similarity and test takers’ race was counter to hypothesis 2.

Research Question 2

To examine whether test taker race moderates the effect of test taker-IP racial similarity on response time, we conducted a GLM with test taker-IP racial similarity and test taker race as independent variables and response times to the four SJT items as the dependent variables. The interaction was statistically significant, Wilk’s Lambda = .94, $F(8,646) = 2.74$, $p < .001$.

As shown in Table 4, Black test takers spent significantly more time responding to each scenario when interacting with racially similar IPs.[8] The amount of time Asian and White test takers spent responding did not differ significantly as a function of their racial similarity to the IPs.

Research Questions 3 and 4

To examine whether IP gender moderates the main effects of test taker-IP racial similarity or the interactive effects of racial

[7] We crossed participant race and IP race to check if results differed depending on the race of the IP and found that for the conflict situation, Black test takers performed better when interacting with a Black IP than when interacting with the White IP. For the helping situation, Black test takers performed better when interacting with a Black IP than when interacting with either the White IP or the Asian IP. However, sample sizes are small for cells crossing Black participants with Asian and White IPs (Ns = 8–9).

[8] When crossing test taker race and IP race, Black test takers spent significantly more time responding to Black IPs than to Asian IPs for all four scenarios. Response time differences for Black IPs relative to White IPs were statistically significant for conflict situation 2 and helping situation 2. Again, sample sizes are small for cells crossing Black participants with Asian and White IPs (Ns = 8–9).

Table 4 Time spent on SJT items as function of test taker-IP racial similarity, study 1

Scenario	Test taker race	Racially similar IP *M* (SD)	Racially dissimilar IP *M* (SD)	*t* test	*d*
Helping 1	Asian	0.92 (0.25)	1.07 (0.40)	$t(35.60) = 1.46$, ns	0.45
	Black	1.37 (0.57)	0.99 (0.40)	$t(28) = -2.15$, $p < .05$	0.77
	White	0.86 (0.25)	0.93 (0.37)	$t(228.06) = 1.73$, ns	0.21
Helping 2	Asian	0.91 (0.26)	0.89 (0.30)	$t(37) = -0.13$, ns	0.04
	Black	1.26 (0.49)	0.81 (0.32)	$t(28) = -3.04$, $p < .01$	1.09
	White	0.81 (0.26)	0.89 (0.35)	$t(261) = 1.83$, ns	0.26
Conflict 1	Asian	1.03 (0.26)	1.01 (0.40)	$t(37) = -0.13$, ns	0.04
	Black	1.25 (0.48)	0.92 (0.32)	$t(28) = -2.24$, $p < .05$	0.80
	White	0.93 (0.35)	1.01 (0.37)	$t(261) = 1.60$, ns	0.22
Conflict 2	Asian	1.00 (0.41)	1.03 (0.47)	$t(37) = 0.20$, ns	0.07
	Black	1.20 (0.54)	0.84 (0.24)	$t(15.44) = -2.44$, $p < .05$	0.85
	White	0.82 (0.31)	0.87 (0.32)	$t(261) = 1.05$, ns	0.14

similarity and test taker race, we reran the earlier analyses with IP gender as an additional independent variable. IP gender did not moderate other variables' effects on performance or response time.[9]

Hypothesis 3

Given that assumed role actors were always White, when IPs were also White, the SJT scenarios depicted no racial diversity. When IPs were either Black or Asian, the level of racial diversity was higher. To examine whether the test taker race moderates the effects of SJT scenario diversity on test takers' perceived fit with the organization, we conducted a GLM with diversity (White IPs vs Black or Asian IPs) and test taker race as independent variables and fit with the organization as the dependent variable. The interaction was not statistically significant, $F(2,329) = 1.04$, MSE = 0.62, ns, and neither was the main effect of diversity, $F(1,329) = 0.00$, ns. Thus, hypothesis 3 was not supported.

Exploratory Analyses

To be thorough, we also examined the interactive effects of test taker-IP racial similarity and test taker race on fit. The interaction was not statistically significant, $F(2,326) = 0.68$, MSE = 0.63, ns, an neither was the main effect of racial similarity, $F(1,326) = 2.09$, ns. Thus, test takers did not perceive different levels of fit with the organization depending on whether they were racially similar to their IPs shown in the scenarios.

An additional approach is to examine whether test takers on average, or specific test taker groups, perceive different levels of fit depending on IPs' minority group membership, as Asian and Black individuals do not have the same social status in American society (Asians are perceived as a "model minority" with high work ethic; Alvarez et al. 2006). We conducted a GLM with IP minority group (Asian vs Black) and test taker race as independent variables and fit with the organization as the dependent variable. The interaction of the IP minority group and test taker race on fit was not statistically significant, $F(2,225) = 1.51$, MSE = 0.63, ns. The main effect of the IP minority group was also not statistically significant, $F(1,225) = 0.06$, ns. Thus, test takers did not perceive significantly different levels of fit with the organization depending on whether they interacted with Asian or Black individuals.

Study 2

Overall, Study 1 indicated that greater attention to how the demography depicted in assessment materials affects test taker responses may be warranted in the development of new tools. A limitation of the study, however, is that all the actors a test taker within a particular condition "interacted with" across scenarios were of a constant race and gender. In practice, when organizations administer video-based SJTs, it is unlikely that actor or avatar demographics will remain constant. Instead, different races and genders would likely be represented across scenarios to convey the organization's diversity. Research on multiple categorization suggests that when forming impressions, perceivers may tend to attend to one category and dampen the activation of others (Bodenhausen and Macrae 1998) and that the relative number and visibility of cues may affect which category dominates the impression

[9] Additional exploratory analyses examined whether test taker-IP gender similarity affected SJT item scores, item response times, or perceived fit with the organization and whether test taker gender interacted with test taker race, test taker-IP racial similarity, or level of diversity represented in the SJT videos. There were no statistically significant effects in Study 1.

Table 5 Study 2 design

	Condition 1		Condition 2		Condition 3		Condition 4		Condition 5		Condition 6	
SJT item	ARA	IP	ARA	IP	ARA	IP	ARA	IP	ARA	IP	ARA	IP
Conflict 1	ARA1	AF1	ARA1	AM1	ARA1	BF1	ARA1	BM1	ARA1	WF1	ARA1	WM1
Helping 1	ARA2	AF2	ARA2	AM2	ARA2	BF2	ARA2	BM2	ARA2	WF2	ARA2	WM2
Conflict 2	ARA3	AM3	ARA3	AF3	ARA3	BM3	ARA3	BF3	ARA3	WM3	ARA3	WF3
Helping 2	ARA4	AM4	ARA4	AF4	ARA4	BM4	ARA4	BF4	ARA4	WM4	ARA4	WF4

Note: *ARA*, assumed role actor, a White female. Assumed role actors indicated with the same number across columns are the same individuals. *IP*, interaction partner; *AF(M)*, Asian female (male); *BF(M)*, Black female (male); *WF(M)*, White female (male). A given actor is identified with the same acronym and number in Study 2 and Study 1 (see Table 1). Scenarios were presented in random order

formation process (Kulik et al. 2007). In Study 2, we kept the race of IPs constant within condition but varied their gender in order to simulate a higher level of demographic variety and examine the effect on the relative salience of the race and gender cues (inferred from cues' associations with the outcomes in the study).

Study 2 Method

Sample

Undergraduates at the same university participated in the second study in exchange for partial course credit (final $N = 304$). Study 1 participants were not allowed to participate in Study 2. Females comprised 54.6% of the sample; 60.5% were White, 23.0% were Asian, and 16.4% were African American/Black. Their average age was 20.12 (SD = 2.80). Finally, they were 32.2% freshmen, 23.0% sophomores, 23.0% juniors, 21.4% seniors, and 0.3% other/non-degree.

Design

We used the same SJT scenarios as in Study 1, but partially rearranged the videos comprising the different conditions. Across and within conditions, the assumed role actor was still a White female (same actress for the same scenario across conditions; different actresses across scenarios within condition). The race of each IP still varied across conditions. Within condition, IPs' race was still constant, but, for this study, their gender varied (two scenarios with male IP, two with female IP). Thus, a given respondent saw four scenarios in all of which the IP was of a constant race but not always of the same gender. Table 5 displays the experimental design.

Measures and Procedure

The procedure for Study 2 paralleled that of Study 1. Participants were asked to imagine themselves as job applicants taking an interpersonal test as an initial step in the process of applying for a technician position and then completed the SJT scenarios and fit measure.

Study 2 Results

Data Cleaning

Data were cleaned in the same manner as in Study 1. We excluded 18 participants who failed the attention check for one or more scenarios by choosing the same option as their most and least likely response to the scenario and 15 additional participants who provided data with multivariate outliers.[10] The final sample size was 304.

Descriptive Statistics

Intercorrelations, means, standard deviations, and reliabilities are provided in Table 6. Male respondents perceived better fit with the organization than female respondents ($r = .14$, $p < .05$).

Hypothesis 1

To test for a positive effect of test taker-IP racial similarity on test taker performance, we conducted a GLM with racial similarity as the independent variable and scores on the four SJT items as the dependent variables. Test takers did not perform better on the scenarios when interacting with racially similar IPs, Wilk's Lambda = .99, $F(4,299) = 1.06$, ns. Thus, similar to Study 1, hypothesis 1 was not supported.

[10] The pattern of significant results did not change when rerunning analyses without excluding multivariate outliers.

Table 6 Descriptives and intercorrelations for Study 2 variables

	M	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1. Scenario diversity	0.65	0.48	–																	
2. IP race	0.52	0.50	--[a]	–																
3. IP gender CS1/HS1	0.52	0.50	0.02	0.00	–															
4. IP gender CS2/HS2	0.48	0.50	− 0.02	0.00	− 1.00	–														
5. Race similarity	0.35	0.48	− 0.43	− 0.05	− 0.04	0.04	–													
6. CS1 score	1.21	1.05	0.02	− 0.06	− 0.03	0.03	0.06	–												
7. CS2 score	0.61	0.85	− 0.01	0.03	− 0.05	0.05	0.06	− 0.03	–											
8. HS1 score	0.71	0.96	− 0.02	− 0.07	0.04	− 0.04	0.09	− 0.03	0.26	–										
9. HS2 score	0.64	0.58	0.02	0.01	0.04	− 0.04	0.00	− 0.03	0.12	0.07	–									
10. Total score	3.16	1.89	0.00	− 0.05	− 0.01	0.01	0.11	0.52	0.61	0.63	0.38	–								
11. CS1 time	1.03	0.38	0.08	− 0.02	− 0.05	0.05	0.01	− 0.03	0.04	− 0.02	0.17	0.04	-							
12. CS2 time	0.92	0.33	0.07	− 0.05	− 0.04	0.04	− 0.02	− 0.02	− 0.05	− 0.03	0.13	− 0.01	0.43	-						
13. HS1 time	1.00	0.36	0.05	− 0.02	− 0.06	0.06	− 0.11	− 0.06	0.03	− 0.05	0.08	− 0.02	0.48	0.39	-					
14. HS2 time	0.89	0.29	0.03	− 0.01	− 0.02	0.02	0.01	0.03	0.08	0.05	0.12	0.12	0.48	0.50	0.45	-				
15. Total time	3.84	1.05	0.08	− 0.03	− 0.05	0.05	− 0.03	− 0.03	0.03	− 0.02	0.17	0.04	0.80	0.75	0.77	0.77	-			
16. Fit	3.44	0.73	− 0.02	− 0.07	0.00	0.00	0.07	0.04	− 0.02	− 0.02	− 0.07	− 0.02	0.02	0.03	0.03	0.09	0.05	0.92		
17. Participant gender	0.45	0.50	0.04	− 0.03	− 0.04	0.04	− 0.13	0.00	− 0.09	− 0.07	− 0.08	− 0.10	0.07	0.08	0.02	0.03	0.06	0.14	-	
18. Participant race	0.61	0.49	− 0.03	− 0.05	0.02	− 0.02	0.03	0.06	− 0.08	0.00	0.02	0.00	− 0.12	− 0.14	− 0.22	− 0.15	− 0.20	− 0.05	0.10	-

Note: $N = 199–304$. *IP*, interaction partner; *CS*, conflict situation; *HS*, helping situation. For scenario diversity, 0 = White interaction partners and 1 = Asian or Black interaction partners. For IP race, 0 = Asian and 1 = Black. For IP gender, 0 = female and 1 = male. For race similarity, 0 = dissimilar to interaction partners and 1 = similar. For participant gender, 0 = female and 1 = male. For participant race, 0 = Asian or Black and 1 = White. Time variables are in minutes.

Correlations in italics are significant at the $p < .05$ level or better. Reliability for the fit measure is shown on the diagnonal.

[a] Correlation cannot be computed because the two variables represent different coding of the race manipulation

Research Question 1

To examine the effects of test taker-IP racial similarity on response time, we conducted a GLM with racial similarity as the independent variable and overall response times to the four SJT items as the dependent variables. As in Study 1, the amount of time test takers spent responding to the scenarios did not differ as a function of test taker-IP racial similarity, Wilk's Lambda = .98, $F(4,299) = 1.61$, ns.

Hypothesis 2

To examine whether test taker race moderates the effects of test taker-IP racial similarity on performance, we conducted a GLM with racial similarity and test taker race (Asian, Black, or White) as independent variables and scores on the four SJT items as the dependent variables. The interaction was not statistically significant, Wilk's Lambda = .99, $F(8,590) = .33$, ns. Thus, hypothesis 2 was not supported. Whereas this is consistent with Study 1 results for White and Asian test takers, who did not perform significantly worse when interacting with racially dissimilar IPs, it is different from Study 1 findings for Black test takers, who had performed significantly worse when interacting with racially dissimilar IPs relative to racially similar IPs.[11]

Research Question 2

To examine whether test taker race moderates the effect of test taker-IP racial similarity on response time, we conducted a GLM with racial similarity and test taker race as independent variables and response times to the four SJT items as the dependent variables. The interaction was not statistically significant, Wilk's Lambda = .98, $F(8,590) = .59$, ns. While this is consistent with Study 1 results for White and Asian test takers, it is different from Study 1 findings for Black test takers, who spent more time responding to the scenarios when interacting with racially similar IPs.[12]

Research Questions 3 and 4

To examine whether IP gender moderates the main effects of test taker-IP racial similarity or the interactive effects of racial similarity and test taker race, we reran the earlier analyses with IP gender as an additional independent variable. Because IP gender varied within condition (see Table 5), we grouped together scenarios that had a common IP gender for the omnibus tests. Consistent with Study 1 results, there were no statistically significant main effects or interactive effects involving IP gender on performance on conflict situation 2 or helping situation 2, and there were no significant effects involving IP gender on response times to the scenarios. However, unique to Study 2, following a significant omnibus test of a racial similarity by IP gender interaction on performance on conflict situation 1 and helping situation 1, Wilk's Lambda = .97, $F(2,291) = 4.27$, $p < .05$, a t test showed that when interacting with a racially similar IP, test takers performed better on conflict situation 1 when the IP was female ($M = 1.52$, SD = .82) as compared with male ($M = 1.06$, SD = 1.16), $t(91.31) = 2.35$, $p < .05$, $d = 0.46$ (medium effect size).

To understand why there may have been score differences, we cross tabulated the response variables with IP gender. The conflict situation involved the assumed role actor noticing a problem on the manufacturing line, with the IP indicating that everything at his/her end is perfect. Test takers selected a keyed "best" response of saying he/she would continue to monitor the situation for their "most likely" choice and a keyed "worst" response of reporting the IP to the supervisor as their "least likely" choice somewhat more frequently after watching a female IP. That is, they performed better with a female than male IP in this scenario; however, the response of waiting to see whether there is a problem can also be viewed as being less willing to listen to a female than male coworker's concern (i.e., differential treatment).

This effect of IP gender was not present for conflict situation 1 interactions with racially *dissimilar* IPs. Racial similarity and IP gender did not interact with helping situation 1 performance.[13]

Hypothesis 3

To examine whether test taker race moderates the effects of SJT scenario racial diversity on test takers' perceived fit with the organization, we conducted a GLM with diversity (White IPs vs Black or Asian IPs) and test taker race as independent variables and fit with the organization as the dependent variable. The interaction was not statistically significant, $F(2,298) = 0.76$, MSE = 0.54, ns, and neither was the main effect of diversity, $F(1,298) = 0.01$, ns. Thus, consistent with Study 1 results, hypothesis 3 was not supported.

[11] There were also no statistically significant differences in performance when crossing test taker race and IP race.

[12] We also crossed test taker race and IP race and found that Black test takers spent more time responding to the two conflict situations when interacting with Asian IPs. For conflict situation 1, the difference was statistically significant relative to White IPs; for conflict situation 2, the difference was statistically significant relative to both White and Black IPs.

[13] Additional exploratory analyses examined whether test taker-IP gender similarity affected SJT item scores, item response times, or perceived fit with the organization and whether test taker gender interacted with test taker race, test taker-IP racial similarity, or level of diversity represented in the SJT videos. Like in Study 1, there were no statistically significant effects.

Exploratory Analyses

We again examined the interactive effects of test taker-IP racial dissimilarity and test taker race on fit. The interaction was not statistically significant, $F(2,298) = 1.05$, MSE = 0.54, ns, and neither was the main effect of racial similarity, $F(1,298) = 1.51$, ns. These results are consistent with those in Study 1.

We also examined whether test takers on average, or specific test taker groups, perceive different levels of fit depending on IPs' minority group membership. The main effect of IP minority group on fit was not statistically significant, $F(1,193) = .59$, MSE = .51, ns, but the interaction of IP minority group and test taker race was, $F(2,193) = 6.46$, $p < .01$. Asian test takers' fit perceptions did not differ significantly depending on whether they interacted with Asian ($M = 3.37$, SD = .75) or Black IPs ($M = 3.56$, SD = .51), $t(45) = -1.01$, ns, $d = .29$ (small effect size). White test takers perceived better fit with the organization when interacting with Asian IPs ($M = 3.57$, SD = .70) than when interacting with Black IPs ($M = 3.16$, SD = .83), $t(116) = 2.88$, $p < .01$, $d = .53$ (medium effect size). Black test takers perceived better fit with the organization when interacting with Black IPs ($M = 3.79$, SD = .62) than when interacting with Asian IPs ($M = 3.30$, SD = .65), $t(32) = -2.24$, $p < .05$, $d = .77$ (large effect size). The results are consistent with Study 1 results for Asian test takers but not for White and Black test takers.

General Discussion

Technology has led to tremendous innovations in the realism and complexity of assessment stimuli (Tippins and Adler 2011). SJTs in particular have moved from written scenarios to the use of video, avatar, and other forms of simulations as question prompts, raising questions as to how demographic information portrayed in these stimuli might influence results. This study contributes to the literature on SJTs in particular and selection assessment tools in general by showcasing how media richness and realism enhancement may have unattended consequences on test taker performance and perceptions. Furthermore, this study extends the increasing body of research on identity contingency cues as influences on applicant attraction (Emerson and Murphy 2014) to an examination of actual assessment content. Additionally, the study provides an illustration of how relational demography research is relevant in the selection context beyond just interviewer-interviewee similarity (Sacco et al. 2003).

Summary of Findings

We conducted two experimental studies with different designs to examine the influence of test taker-interaction partner race similarity and interaction partner gender on responses to SJT scenarios and reactions. Across the two studies, we did not find a main effect of racial similarity on performance (H1) or time spent (RQ1). With regard to interaction effects (H2 and RQ2), we found similar results across studies for White and Asian test takers but not for Blacks. We also found largely similar effects across the two studies for gender (RQ3 and 4), except in one scenario. In addition to these similar findings, there were some differences across studies. First, we found several significant effects of dissimilarity in race on responses to SJT scenarios. Black test takers scored better when interacting with racially similar individuals (in Study 1 but not Study 2) and spent more time responding to the SJT scenarios when interacting with similar individuals (in Study 1 but not Study 2). For one of the conflict scenarios, test takers performed better when interacting with a racially similar *female* IP than a racially similar *male* IP (in Study 2 but not Study 1). White test takers perceived a better fit with the organization when interacting with Asian versus Black individuals and Black test takers perceived better fit when interacting with Black versus Asian individuals (in Study 2 but not Study 1). Note that the significant effects for performance and response time were medium to large in magnitude ($d = .46$ or greater), indicating that cumulative effects across an entire assessment might make a considerable difference in scores, and ultimately in which individuals are hired.

Theoretical Implications

Our findings contribute to research in areas of selection testing, applicant recruitment, and interracial interactions. In terms of selection assessment research, our findings indicate the need to extend our work and conceptualization of the effects of technology on constructs assessed. As assessment developers move toward richer media such as videos, animations, and even virtual reality within assessment tools, there is a need for theoretical guidance as to aspects of richer media that may influence performance, especially to the extent to which the constructs assessed might be degraded. While we focused on the specific issue of the richer demographic information that richer assessment stimuli provide, newer tools also provide greater context information (i.e., visual images of a specific workplace) that are also attended to by test takers. Lievens and Sackett's (2017) modular approach to personnel selection procedures provides an excellent starting point for this discussion, as they specifically note stimulus format (e.g., textual, pictorial) and level of contextualization in stimuli as important variants in building assessments.

Whereas researchers have tended to examine the influence of identity contingency cues present in traditional recruitment materials (e.g., ads, brochures; Avery et al. 2004; Purdie-Vaughns et al. 2008) on reactions, we extend this research to assessment stimuli and additional outcomes of interest (e.g., test performance). Furthermore, scenario-based selection assessments have not been previously examined through the lens of ingroup/outgroup interactions. Given the differences in findings across studies, results seem to suggest that the effects of demographic cues in assessments on test takers may be complex and warrant further research.

Interestingly, Study 1 SJT scenario performance and response time-related findings and Study 2 fit-related findings for Black respondents are more consistent with research on diversity cues and processing of recruitment information than they are with research on interracial interactions, highlighting that assessment tools can serve double duty as selection and recruitment tools. For minorities, the presence of diversity cues in recruitment materials can signal a diverse organizational climate increasing their attraction to the organization and thereby motivating them to process presented information more thoroughly (Walker et al. 2012). As further discussed shortly, we recommend investigating the potential explanatory mechanisms.

Although Asians are considered a minority group, the fact that Asian test takers in the current studies were generally not influenced by racial dissimilarity with their interaction partners is consistent with the notion that Asians' experiences as a minority group may be different from those of other minorities in the USA (Unzueta and Binning 2010). In contrast to Blacks and Latinos, Asians are often viewed as model minorities (Alvarez et al. 2006; Shelton et al. 2010). Furthermore, the presence of Asians in an organization may not necessarily enhance the perceived diversity of that organization (Unzueta and Binning 2010), though we might expect regional differences in such effects. The fact that White test takers in Study 2 reported better perceived fit with an organization that included Asian relative to Black individuals seems to imply that these test takers viewed Asians as more similar to themselves.

Even though the same scenarios were used in both studies, findings were not consistent. While we cannot rule out the potential that findings are due to somewhat underpowered tests of interactions when dealing with minority samples,[14] we think the design difference—varying interaction partner gender between subjects in Study 1 but within subjects in Study 2—was a key contributor. Research suggests that for ease of information processing, individuals may attend to one category when forming impressions, and context helps to determine which category will be dominant (Bodenhausen and Macrae 1998; Kulik et al. 2007). In Study 1, with interaction partners' gender held constant within condition, their race may have been more salient; in Study 2, varying interaction partners' gender within condition may have made their race less salient. Furthermore, research on intersectionality (Cole 2009; Crenshaw 1989/1993, 1991) suggests attending to multiple social categories simultaneously when considering how an individual will be perceived; for example, how a Black male might be viewed in an interracial interaction may be different than how a Black female in the same interaction might be considered (Sesko and Biernat 2010). Thus, we would urge future research on simulated interracial interactions in assessment contexts to employ designs that allow considering race, gender, and their intersection to build on current findings.

Practical Implications

While our findings were not completely in line with our hypotheses, our results did show that even in a simulated interaction of short duration, demographic dissimilarity may affect performance, time spent, and reactions. Indeed, while not replicated in Study 2, the finding of lower performance of Black test takers when interacting with dissimilar others in Study 1 highlights the importance of considering how a workplace scenario (e.g., a conflict with a coworker) might be interpreted as requiring different actions when it involves interaction with an individual of a different race. If individuals do respond differently to scenarios based on their own relative to an actor's demographics, these differences may translate into group performance differences on the assessment. Group differences are important in selection contexts (Ployhart and Holtz 2008) and test developers may want to consider how to minimize possible negative effects of demographic cues in videos or animations on adverse impact statistics.

One strategy may be to ask sensitivity and fairness experts to review drafted scenarios and planned demographic cues before the media is created. While it may seem apparent that for an inclusive workplace, one would not want to hire applicants who would react differently to coworkers based on their demographic background, research does show that such differences exist (e.g., interracial interactions can produce anxiety [Stephan and Stephan 1985], people tend to help women more than men [Eagly and Crowley 1986]). Awareness of this in assessment design can ensure thoughtful attention to how individuals of different races and genders are portrayed in these scenarios, as well as how test takers might perceive the same situation if different social categories (and potentially associated stereotypes) are activated.

Many organizations strive for demographic diversity in their recruitment and assessment materials so as to enhance applicant reactions and attract diverse applicants. Ensuring diversity across assessment stimuli (e.g., representing different races and genders) may also be an effective strategy for minimizing group performance differences to the extent that any performance gains (decrements) associated with individual stimuli featuring similar (dissimilar) others may average

[14] For interactions involving six groups (e.g., race similarity × test taker race), a sample size of approximately 35 per group is recommended to detect medium effect sizes at a .05 significance level with a power of .80 (Cohen 1992). After data cleaning, our sample sizes of minority test takers ranged from 11 to 18 in Study 1 and 15 to 36 in Study 2.

out across a longer set of diverse stimuli. Relatedly, for SJTs, having multiple situations of a particular type (e.g., conflict, helping), but where demographic cues are varied (e.g., in one situation a Black female coworker needs help, in another an Asian male), may minimize adverse impact concerns while also enhancing assessment reliability. The efficacy of these recommendations should be tested via further research, as the current research did not examine the effectiveness of these strategies.

Limitations and Future Directions

A number of limitations in the current research should be noted. First, lab studies like the current ones can have limitations in that results may not always generalize to other settings. Watching interpersonal interactions likely allows individuals more cognitive distance from the interaction partners shown than would actual interactions with these individuals. Imagined interracial interactions may not elicit discomfort to the same extent that actual interracial interactions would. However, actual interactions are rare in screening devices (i.e., more and more assessment centers have gone virtual, using video or animated stimuli) and would be expected to become increasingly rare with the greater use of augmented reality and virtual reality in multimedia assessments (Boyce et al. 2013).

Also related to generalizability, although we asked participants to imagine themselves in a selection situation, the current context was low stakes. Higher-stakes settings may be associated with higher levels of anxiety than may have been experienced by the average participant in the current research. In a related vein, college students may hold more overtly positive racial and gender-related attitudes relative to the general adult population (given educational institutions' efforts to increase demographic diversity and expose students to diverse points of view; Worthington et al. 2008). From the perspective that interactions with outgroup members are likely to be anxiety provoking, especially when self-presentational concerns are salient and for those who lack relevant intergroup experience, demographic cue effects in assessment stimuli may actually be more common when examined in high-stakes settings and with working adult samples.

However, given our hypotheses that demographic cues in the assessment would influence test takers' performance and reactions, we believe that this research was more appropriate to conduct in a lab setting than with job applicants in a high-stakes, selection setting. It would not be ethical to treat job applicants inconsistently in a way that we expected would influence their test performance and reactions, thereby affecting certain groups' chances of obtaining employment or perceived fit with the organization. Furthermore, organizations may not readily agree to allow researchers to manipulate their selection materials in ways that may misrepresent their levels of organizational diversity or lead to legal challenges.

Second, we did not examine explanatory mechanisms for demographic cue effects on respondents. Research points to a number of mediators (e.g., anxiety, uncertainty, distraction, attraction) and moderators (e.g., amount of experience interacting with outgroup members) of potential racial dissimilarity effects (e.g., Plant and Butz 2006; Richeson and Shelton 2003; Stephan and Stephan 1985). Attitudes (e.g., held racial or gender stereotypes, stigma consciousness, identity centrality) and motivations (e.g., to respond to outgroup members without prejudice, to be culturally sensitive, to avoid confirming a stereotype about one's group) may also play a role (e.g., Babbitt 2013; Dovidio and Gaertner 2004; Pinel 1999; Trawalter et al. 2009). Given the desire to create a simulated selection experience, we did not focus on assessing these established connections in this particular study; however, additional lab-based, experimental research to examine potential explanatory mechanisms to better understand how test takers interpret and react to demographic cues in assessment stimuli as well as which test takers are more likely to be affected by such cues (e.g., examining the effects of test takers' age or experience) is warranted. Notably, research in related domains, such as investigations of the mediators of stereotype threat effects on performance (e.g., Schmader and Johns 2003) can provide guidance on addressing similar questions in the assessment context.

Third, because of the high costs involved in casting actors and professionally filming multiple versions of each SJT scenario, we examined only a small number of SJT situations in the current research. We made an effort to include common types of SJT scenarios (conflict, helping) to enhance the relevance of our findings for other video-based SJTs. However, it would be important to examine generalizability of current findings to other types of scenarios within the conflict and helping domains, to scenarios within other behavioral domains of interest in candidate assessment, and to longer assessments. We also note that not all SJT items involve interpersonal contexts and thus demographic cues (e.g., a procedural knowledge type of question), although many do. Cognitive depletion, which may mediate effects of racial dissimilarity, is more likely with longer assessments; on the other hand, as we suggested earlier, demographic cue effects present in certain situations potentially wash out over longer assessments. We would also recommend examining how the content of assessment scenarios (e.g., the individual differences scenarios are meant to elicit, the attitudes/stereotypes scenarios may activate) moderates demographic cue effects on individuals' performance and reactions. To explore this question, an assumed role actor could be shown interacting with the same interaction partner across scenarios of varying content. Furthermore, if performance differences are found in future research, examining effects on validity would be another important consideration to pursue.

Fourth, to keep a complex design from becoming even more complex, we kept the assumed role actor demographics constant and therefore did not examine the effects of assumed role actor demographics on performance and reactions. Research on avatars has begun to address implications of choosing or being assigned avatars that differ from oneself in gender and ethnicity (e.g., Ratan and Sah 2015). Incorporating concepts such as avatar (or video assumed role actor) self-relevance into future research may be fruitful.

Finally, assessment stimuli that present actors or avatars may vary in the format used to elicit responses from test takers. In the current research, that response format involved test takers choosing a most and least likely response from a set of options. Research indicates that interracial interactions with greater structure are associated with better outcomes for those involved as compared with more free-form interactions, arguably because structure cues individuals on the types of responses that may be appropriate and reduces their anxiety (Toosi et al. 2012; Trawalter et al. 2009). As such, individuals may experience greater anxiety and performance decrements interacting with outgroup members relative to ingroup members when they have to construct their own responses, as opposed to selecting a response option. On the other hand, a constructed response SJT format may be associated with smaller subgroup score differences than a multiple choice SJT format due to a lower cognitive load (based on a study with Belgian natives and immigrants; Lievens et al. 2019). Future research can examine this question.

Conclusion

Technology has enabled selection assessment developers to create more realistic simulations. Our studies indicate that attention needs to be paid toward considering how social identities are portrayed in these newer assessment forms, as this information may affect performance and applicant reactions. As organizations seek to reduce adverse impact and to attract diverse workforces, subtle cues such as demographics in assessments deserve greater consideration in our research and practice.

Acknowledgments The authors wish to thank Jessie Olien, Abraham Kim, and Julie Klopp for their assistance with data collection for the current studies. The authors also thank Neal Schmitt, Danielle King, Dia Chatterjee, and two anonymous reviewers for their comments on an earlier version of this paper.

Compliance with Ethical Standards This research was approved by the Institutional Review Board at Michigan State University.

Conflict of Interest The authors declare that they have no conflict of interest.

References

Alvarez, A. N., Juang, L., & Liang, C. T. H. (2006). Asian Americans and racism: When bad things happen to “model minorities”. *Cultural Diversity and Ethnic Minority Psychology, 12*(3), 477–492. https://doi.org/10.1037/1099-9809.12.3.477.

Apfelbaum, E. P., Sommers, S. R., & Norton, M. I. (2008). Seeing race and seeming racist? Evaluating strategic colorblindness in social interaction. *Journal of Personality and Social Psychology, 95*(4), 918–932. https://doi.org/10.1037/a0011990.

Avery, D. R. (2003). Reactions to diversity in recruitment advertising: Are differences black and white? *Journal of Applied Psychology, 88*(4), 672–679. https://doi.org/10.1037/0021-9010.88.4.672.

Avery, D. R., & McKay, P. F. (2006). Target practice: An organizational impression management approach to attracting minority and female job applicants. *Personnel Psychology, 59*(1), 157–187. https://doi.org/10.1111/j.1744-6570.2006.00807.x.

Avery, D. R., Hernandez, M., & Hebl, M. R. (2004). Who’s watching the race? Racial salience in recruitment advertising. *Journal of Applied Social Psychology, 34*(1), 146–161. https://doi.org/10.1111/j.1559-1816.2004.tb02541.x.

Babbitt, L. G. (2013). An intersectional approach to Black/White interracial interactions: The roles of gender and sexual orientation. *Sex Roles, 68*(11–12), 791–802. https://doi.org/10.1007/s11199-011-0104-4.

Balliet, D., Li, N. P., Macfarlan, S. J., & Van Vugt, M. (2011). Sex differences in cooperation: A meta-analytic review of social dilemmas. *Psychological Bulletin, 137*(6), 881–909. https://doi.org/10.1037/a0025354.

Bauer, T. N., & Truxillo, D. M. (2006). Applicant reactions to situational judgment tests: Research and related practical issues. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 233–251). Mahwah, NJ: Erlbaum.

Bodenhausen, G. V., & Macrae, C. N. (1998). Stereotype activation and inhibition. In R. S. Wyer Jr. (Ed.), *Stereotype activation and inhibition: Advances in social cognition* (pp. 1–52). Hillsdale, NJ: Erlbaum.

Boyce, A.S, Corbet, CE & Adler, S (2013). Simulations in the selection context: Considerations, challenges, and opportunities. In M. Fetzer & K. Tuzinski (Eds.). Simulations for personnel selection (pp17–41). New York, NY: Springer.

Campion, M. C., Ployhart, R. E., & MacKenzie Jr., W. I. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. *Human Performance, 27*(4), 283–310. https://doi.org/10.1080/08959285.2014.929693.

Carli, L. L. (1989). Gender differences in interaction style and influence. *Journal of Personality and Social Psychology, 56*(4), 565–576.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82*(1), 143–159. https://doi.org/10.1037/0021-9010.82.1.143.

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance, 15*(3), 233–254. https://doi.org/10.1207/S15327043HUP1503_01.

Chaney, K. E., & Sanchez, D. T. (2018). Gender-inclusive bathrooms signal fairness across identity dimensions. *Social Psychological and Personality Science, 9*(2), 245–253. https://doi.org/10.1177/1948550617737601.

Chaney, K. E., Sanchez, D. T., & Remedios, J. D. (2016). Organizational identity safety cue transfers. *Personality and Social Psychology Bulletin, 42*(11), 1564–1576. https://doi.org/10.1177/0146167216665096.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63*(1), 83–117. https://doi.org/10.1111/j.1744-6570.2009.01163.x.

Clark, R., Anderson, N. B., Clark, V. R., & Williams, D. R. (1999). Racism as a stressor for African Americans: A biopsychosocial model. *American Psychologist, 54*(10), 805–816. https://doi.org/10.1037/0003-066X.54.10.805.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Cole, E. R. (2009). Intersectionality and research in psychology. *American Psychologist, 64*(3), 170–180. https://doi.org/10.1037/a0014564.

Crenshaw, K. (1989/1993). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist policies. In D. K. Weisbert (Ed.), Feminist legal theory: Foundations (pp. 383–395). Philadelphia, PA: Temple University Press. (Original work published 1989).

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review, 43*(6), 1241–1299. https://doi.org/10.2307/1229039.

de Meijer, L. A. L., Born, M. P., van Zielst, J., & van der Molen, H. T. (2010). Construct-driven development of a video-based situational judgment test for integrity: A study in a multi-ethnic police setting. *European Psychologist, 15*, 229–236.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*(1), 5–18. https://doi.org/10.1037/0022-3514.56.1.5.

Doerr, C., Plant, E. A., Kunstman, J. W., & Buck, D. (2011). Interactions in Black and White: Racial differences and similarities in response to interracial interactions. *Group Processes & Intergroup Relations, 14*(1), 31–43. https://doi.org/10.1177/1368430210375250.

Dovidio, J. F., & Gaertner, S. L. (2004). Aversive racism. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 36, pp. 1–52). San Diego, CA: Elsevier Academic Press.

Eagly, A. H., & Crowley, M. (1986). Gender and helping behavior: A meta-analytic review of the social psychological literature. *Psychological Bulletin, 100*(3), 283–308. https://doi.org/10.1037/0033-2909.100.3.283.

Emerson, K. T. U., & Murphy, M. C. (2014). Identity threat at work: How social identity threat and situational cues contribute to racial and ethnic disparities in the workplace. *Cultural Diversity and Ethnic Minority Psychology, 20*(4), 508–520. https://doi.org/10.1037/a0035403.

Goldberg, C., Riordan, C. M., & Zhang, L. (2008). Employees' perceptions of their leaders: Is being similar always better? *Group & Organization Management, 33*(3), 330–355. https://doi.org/10.1177/1059601108318232.

Graves, L. M., & Elsass, P. M. (2005). Sex and sex dissimilarity effects in ongoing teams: Some surprising findings. *Human Relations, 58*(2), 191–221. https://doi.org/10.1177/0018726705052181.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Lawrence Erlbaum.

Hawkes, B., Cek, I., & Handler, C. (2018). The gamification of employee selection tools: An exploration of viability, utility, and future directions. In J. C. Scott, D. Bartram, & D. H. Reynolds (Eds.), *Next generation technology-enhanced assessment: Global perspectives on occupational and workplace testing* (pp. 288–316). Cambridge, UK: Cambridge University Press.

Hussain, Z., & Griffiths, M. D. (2008). Gender swapping and socializing in cyberspace: An exploratory study. *Cyberpsychology & Behavior, 11*(1), 47–53. https://doi.org/10.1089/cpb.2007.0020.

Ito, T. A., & Urland, G. R. (2003). Race and gender on the brain: Electrocortical measures of attention to the race and gender of multiply categorizable individuals. *Journal of Personality and Social Psychology, 85*(4), 616–626. https://doi.org/10.1037/0022-3514.85.4.616.

Kirchmeyer, C. (1995). Demographic similarity to the work group: A longitudinal study of managers at the early career stage. *Journal of Organizational Behavior, 16*(1), 67–83. https://doi.org/10.1002/job.4030160109.

Kulik, C. T., Roberson, L., & Perry, E. L. (2007). The multiple-category problem: Category activation and inhibition in the hiring process. *Academy of Management Review, 32*(2), 529–548. https://doi.org/10.5465/amr.2007.24351855.

Lau, D. C., Lam, L. W., & Salamon, S. D. (2008). The impact of relational demographics on perceived managerial trustworthiness: Similarity or norms? *The Journal of Social Psychology, 148*(2), 187–209. https://doi.org/10.3200/SOCP.148.2.187-209.

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

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*(1), 43–66. https://doi.org/10.1037/apl0000160.

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

MacCann, C., Lievens, F., Libbrecht, N., & Roberts, R. D. (2016). Differences between multimedia and text-based assessments of emotion management: An exploration with the multimedia emotion management assessment (MEMA). *Cognition and Emotion, 30*, 1317–1331.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Science of India, 12*, 49-55.

Major, B., Quinton, W. J., & McCoy, S. K. (2002). Antecedents and consequences of attributions to discrimination: Theoretical and empirical advances. *Advances in Experimental Social Psychology, 34*, 251–330. https://doi.org/10.1016/S0065-2601(02)80007-7.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48*, 61–83. https://doi.org/10.1016/j.jrp.2013.09.008.

Martey, R. M., Stromer-Galley, J., Banks, J., Wu, J., & Consalvo, M. (2014). The strategic female: Gender-switching and player behavior in online games. *Information, Communication & Society, 17*(3), 286–300. https://doi.org/10.1080/1369118X.2013.874493.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455. https://doi.org/10.1037/a0028085.

Mendoza-Denton, R., Downey, G., Purdie, V. I., Davis, A., & Pietrzak, J. (2002). Sensitivity to status-based rejection: Implications for African American students' college experience. *Journal of Personality and Social Psychology, 83*(4), 896–918. https://doi.org/10.1037//0022-3514.83.4.896.

Montepare, J. M., & Opeyo, A. (2002). The relative salience of physiognomic cues in differentiating faces: A methodological tool. *Journal of Nonverbal Behavior, 26*(1), 43–59. https://doi.org/10.1023/A:1014470520593.

Moran, T. P. (2016). Anxiety and working memory capacity: A meta-analysis and narrative review. *Psychological Bulletin, 142*(8), 831–864. https://doi.org/10.1037/bul0000051.

Murphy, M. C., & Taylor, V. J. (2012). The role of situational cues in signaling and maintaining stereotype threat. In M. Inzlicht & T. Schmader (Eds.), *Stereotype threat: Theory, process, and* *application; stereotype threat: Theory, process, and application* (pp. 17–33). New York, NY: Oxford University Press.

Murphy, M. C., Steele, C. M., & Gross, J. J. (2007). Signaling threat: How situational cues affect women in math, science, and engineering settings. *Psychological Science, 18*(10), 879–885. https://doi.org/10.1111/j.1467-9280.2007.01995.x.

O'Connell, M., Lawrence, A., & Kinney, T. (2013). Show me you can do it: The use of interactive simulations in manufacturing settings. In M. Fetzer & K. Tuzinski (Eds.), *Simulations for personnel selection* (pp. 157–185). New York, NY: Springer.

Olson-Buchanan, J. B., Drasgow, F., Moberg, P. J., Mead, A. D., Keenan, P. A., & Donovan, M. A. (1998). Interactive video assessment of conflict resolution skills. *Personnel Psychology, 51*(1), 1–24. https://doi.org/10.1111/j.1744-6570.1998.tb00714.x.

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology, 89*(2), 187–208. https://doi.org/10.1037/0021-9010.89.2.187.

Perkins, L. A., Thomas, K. M., & Taylor, G. A. (2000). Advertising and recruitment: Marketing to minorities. *Psychology & Marketing, 17*(3), 235–255. https://doi.org/10.1002/(SICI)1520-6793(200003)17:3<235::AID-MAR3>3.0.CO;2-#.

Pinel, E. C. (1999). Stigma consciousness: The psychological legacy of social stereotypes. *Journal of Personality and Social Psychology, 76*(1), 114–128. https://doi.org/10.1037/0022-3514.76.1.114.

Plant, E. A. (2004). Responses to interracial interactions over time. *Personality and Social Psychology Bulletin, 30*(11), 1458–1471. https://doi.org/10.1177/0146167204264244.

Plant, E. A., & Butz, D. A. (2006). The causes and consequences of an avoidance-focus for interracial interactions. *Personality and Social Psychology Bulletin, 32*(6), 833–846. https://doi.org/10.1177/0146167206287182.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61*(1), 153–172. https://doi.org/10.1111/j.1744-6570.2008.00109.x.

Podsakoff, P. M., MacKenzie, S. B., Paine, J. B., & Bachrach, D. G. (2000). Organizational citizenship behaviors: A critical review of the theoretical and empirical literature and suggestions for future research. *Journal of Management, 26*(3), 513–563. https://doi.org/10.1177/014920630002600307.

Purdie-Vaughns, V., Steele, C. M., Davies, P. G., Ditlmann, R., & Crosby, J. R. (2008). Social identity contingencies: How diversity cues signal threat or safety for African Americans in mainstream institutions. *Journal of Personality and Social Psychology, 94*(4), 615–630. https://doi.org/10.1037/0022-3514.94.4.615.

Randall, J. G., Oswald, F. L., & Beier, M. E. (2014). Mind-wandering, cognition, and performance: A theory-driven meta-analysis of attention regulation. *Psychological Bulletin, 140*(6), 1411–1431. https://doi.org/10.1037/a0037428.

Ratan, R., & Sah, Y. J. (2015). Leveling up on stereotype threat: The role of avatar customization and avatar embodiment. *Computers in Human Behavior, 50*, 367–374. https://doi.org/10.1016/j.chb.2015.04.010.

Richeson, J. A., & Shelton, J. N. (2003). When prejudice does not pay: Effects of interracial contact on executive function. *Psychological Science, 14*(3), 287–290. https://doi.org/10.1111/1467-9280.03437.

Richeson, J. A., & Shelton, J. N. (2004). Brief report: Thin slices of racial bias. *Journal of Nonverbal Behavior, 29*(1), 75–86. https://doi.org/10.1007/s10919-004-0890-2.

Richeson, J. A., & Trawalter, S. (2005). Why do interracial interactions impair executive function? A resource depletion account. *Journal of Personality and Social Psychology, 88*(6), 934–947. https://doi.org/10.1037/0022-3514.88.6.934.

Richeson, J. A., Trawalter, S., & Shelton, J. N. (2005). Racial minorities' implicit racial attitudes and depletion of executive function after interracial interactions. *Social Cognition, 23*(4), 336–352. https://doi.org/10.1521/soco.2005.23.4.336.

Rothbart, M., & John, O. P. (1985). Social categorization and behavioral episodes: A cognitive analysis of the effects of intergroup contact. *Journal of Social Issues, 41*(3), 81–104. https://doi.org/10.1111/j.1540-4560.1985.tb01130.x.

Russell, A., & Owens, L. (1999). Peer estimates of school-aged boys' and girls' aggression to same-and cross-sex targets. *Social Development, 8*(3), 364–379. https://doi.org/10.1111/1467-9507.00101.

Ryan, C. S., Hunt, J. S., Weible, J. A., Peterson, C. R., & Casas, J. F. (2007). Multicultural and colorblind ideology, stereotypes, and ethnocentrism among Black and White Americans. *Group Processes & Intergroup Relations, 10*(4), 617–637. https://doi.org/10.1177/1368430207084105.

Ryan, A. M., Inceoglu, I., Bartram, D., Golubovich, J., Grand, J. A., Reeder, M., Derous, E., Nikolaou, I., & Yao, X. (2015). Trends in testing: Highlights of a global survey. In I. Nikolaou & J. K. Oostrom (Eds.), *Employee recruitment, selection, and assessment: Contemporary issues for theory and practice* (pp. 136–153). New York, NY: Psychology Press.

Sacco, J. M., Scheu, C. R., Ryan, A. M., & Schmitt, N. (2003). An investigation of race and sex similarity effects in interviews: A multilevel approach to relational demography. *Journal of Applied Psychology, 88*(5), 852–865. https://doi.org/10.1037/0021-9010.88.5.852.

Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology, 85*(3), 440–452. https://doi.org/10.1037/0022-3514.85.3.440.

Sesko, A. K., & Biernat, M. (2010). Prototypes of race and gender: The invisibility of Black women. *Journal of Experimental Social Psychology, 46*(2), 356–360. https://doi.org/10.1016/j.jesp.2009.10.016.

Shelton, J. N. (2003). Interpersonal concerns in social encounters between majority and minority group members. *Group Processes & Intergroup Relations, 6*(2), 171–185. https://doi.org/10.1177/1368430203006002003.

Shelton, J. N., & Richeson, J. A. (2006). Interracial interactions: A relational approach. *Advances in Experimental Social Psychology, 38*, 121–181. https://doi.org/10.1016/S0065-2601(06)38003-3.

Shelton, J. N., Richeson, J. A., & Salvatore, J. (2005). Expecting to be the target of prejudice: Implications for interethnic interactions. *Personality and Social Psychology Bulletin, 31*(9), 1189–1202. https://doi.org/10.1177/0146167205274894.

Shelton, J. N., West, T. V., & Trail, T. E. (2010). Concerns about appearing prejudiced: Implications for anxiety during daily interracial interactions. *Group Processes & Intergroup Relations, 13*(3), 329–344. https://doi.org/10.1177/1368430209344869.

Shute, R., & Charlton, K. (2006). Anger or compromise? Adolescents' conflict resolution strategies in relation to gender and type of peer relationship. *International Journal of Adolescence and Youth, 13*(1–2), 55–69. https://doi.org/10.1080/02673843.2006.9747966.

Stephan, W. G., & Stephan, C. W. (1985). Intergroup anxiety. *Journal of Social Issues, 41*(3), 157–175. https://doi.org/10.1111/j.1540-4560.1985.tb01134.x.

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–48). Monterey, CA: Brooks/Cole.

Tippins, N. T., & Adler, S. (2011). *Technology-enhanced assessment of talent*. San Francisco, CA: Jossey-Bass.

Tjosvold, D. (1998). Cooperative and competitive goal approach to conflict: Accomplishments and challenges. *Applied Psychology. An*

International Review, 47(3), 285–313. https://doi.org/10.1111/j.1464-0597.1998.tb00025.x.

Toosi, N. R., Babbitt, L. G., Ambady, N., & Sommers, S. R. (2012). Dyadic interracial interactions: A meta-analysis. *Psychological Bulletin, 138*(1), 1–27. https://doi.org/10.1037/a0025767.

Trawalter, S., Richeson, J. A., & Shelton, J. N. (2009). Predicting behavior during interracial interactions: A stress and coping approach. *Personality and Social Psychology Review, 13*(4), 243–268. https://doi.org/10.1177/1088868309345850.

Tsui, A. S., Egan, T. D., & O'Reilly III, C. A. (1992). Being different: Relational demography and organizational attachment. *Administrative Science Quarterly, 37*, 549–579.

Unzueta, M. M., & Binning, K. R. (2010). Which racial groups are associated with diversity? *Cultural Diversity and Ethnic Minority Psychology, 16*(3), 443–446. https://doi.org/10.1037/a0019723.

Vorauer, J. D. (2006). An information search model of evaluative concerns in intergroup interaction. *Psychological Review, 113*(4), 862–886. https://doi.org/10.1037/0033-295X.113.4.862.

Walker, H. J., Feild, H. S., Bernerth, J. B., & Becton, J. B. (2012). Diversity cues on recruitment websites: Investigating the effects on job seekers' information processing. *Journal of Applied Psychology, 97*(1), 214–224. https://doi.org/10.1037/a0025847.

Weekley, J. A., & Jones, C. (1997). Video-based situational testing. *Personnel Psychology, 50*(1), 25–49. https://doi.org/10.1111/j.1744-6570.1997.tb00899.x.

West, T. V., & Dovidio, J. F. (2013). Intergroup contact across time: Beyond initial contact. In G. Hodson & M. Hewstone (Eds.), *Advances in intergroup contact* (pp. 152–175). New York, NY: Psychology Press.

Worthington, R. L., Navarro, R. L., Loewy, M., & Hart, J. (2008). Color-blind racial attitudes, social dominance orientation, racial-ethnic group membership and college students' perceptions of campus climate. *Journal of Diversity in Higher Education, 1*(1), 8–19. https://doi.org/10.1037/1938-8926.1.1.8.

Wout, D. A., Murphy, M. C., & Barnett, S. (2014). When having Black friends isn't enough: How threat cues undermine safety cues in friendship formation. *Social Psychological and Personality Science, 5*(7), 844–851. https://doi.org/10.1177/1948550614535820.

Yee, N., & Bailenson, J. N. (2007). The Proteus effect: The effect of transformed self-representation on behavior. *Human Communication Research, 33*(3), 271–290. https://doi.org/10.1111/j.1468-2958.2007.00299.x.