Running Head: SENSITIVITY REVIEW PRACTICES FOR TEST DEVELOPMENT

The Detection and Influence of Problematic Item Content in Ability Tests:

An Examination of Sensitivity Review Practices for Personnel Selection Test Development

James A. Grand

The University of Akron

Juliya Golubovich, Ann Marie Ryan, and Neal Schmitt

Michigan State University

Citation:

Grand, J.A., Golubovich, J., Ryan, A.M., & Schmitt, N. (2013). The detection and influence of problematic item content in ability tests: An examination of sensitivity review practices for personnel selection test development. *Organizational Behavior and Human Decision Processes*, *121*, 158-173.

Abstract

In organizational and educational practices, sensitivity reviews are commonly advocated techniques for reducing test bias and enhancing fairness. In the present paper, results from two studies are reported which investigate how effective individuals are at detecting problematic test content and the influence such content has on important testing outcomes. In Study 1, signal detection analyses are used to examine the role of individual differences in the identification of insensitive test items, while Study 2 investigates the extent to which insensitivity differentially influences item performance and reactions. Results revealed small but significant differences in the overall accuracy and response tendencies of student test reviewers on the basis of demographics and key individual differences variables. Contrary to predictions however, problematic items did not exhibit differential item functioning across sex nor did their presence engender negative test taker reactions. Implications and suggestions for future research and sensitivity review practices are discussed.

Keywords: sensitivity review, fairness review, test review, signal detection analysis, test bias, test development, differential item functioning, selection, assessment

Fairness in testing has been a prominent concern of selection specialists for several decades. While organizational psychologists have given considerable research attention to the general topic of adverse impact and test discrimination (cf., Sackett, Schmitt, Ellingson & Kabin, 2001), Ployhart and Holtz (2008) note that evidence for the effectiveness of many methods of improving the fairness of evaluative measures is anecdotal and lacking in rigorous empirical examinations. This paper examines one such fairness evaluation technique—the *sensitivity review*, also referred to as a bias review or fairness review (ETS, 2009; Ramsey, 1993). The primary purpose of a sensitivity review is to remove test content that might prevent or distract test takers from responding in ways that allow for correct inferences about their standing on the measured construct (Zieky, 2006). Some test developers may also commission sensitivity reviews in the belief that they improve an evaluative assessment's psychometric quality or in efforts to proactively improve an evaluation's legal defensibility (McPhail, 2010). Regardless of their intended benefit, sensitivity reviews are primarily conducted to ensure that the test: 1) reflects the cultural background of both majority and minority test takers; 2) is devoid of content considered sexist, racist, offensive, or inappropriate; and 3) has an item format that is accessible to and non-discriminatory towards subgroups of test-takers (ETS, 2002).

Recruited reviewers commonly evaluate the degree to which test items conform to sensitivity guidelines established by the test developer and, if an item does not appear to meet these standards, recommend its exclusion or revision (Johnstone, Thompson, Bottsford-Miller, & Thurlow, 2008; Reckase, 1996). More generally then, sensitivity reviews reflect an evaluative process in which individuals make judgments about the extent to which a stimulus material meets and/or exceeds some subjectively determined criteria that qualifies an item as problematic. For example, sensitivity guidelines often indicate that items with women portrayed in only sextyped roles, terminology that could be differentially familiar across groups (e.g., sports references), insensitive labels (e.g., crippled) and non-inclusive language (e.g., mankind), or graphics that lack diversity or contain stereotypic depictions qualify as problematic. Sensitivity reviewers evaluate a large set of items and provide their subjective judgment on whether any of them possess such problematic content or could otherwise be perceived by test takers as unfair.

While a number of resources elaborate upon guidelines for categorizing problematic content, relatively little attention has been given to the nature and outcomes of either reviewers' or test takers' evaluation of and experience with problematic items. Such information, however, could have important implications for many practical questions surrounding the sensitivity review process, such as who should serve as reviewers, how successful reviews are at removing problematic items, and how problematic content impacts test-taker performance and reactions (Ployhart & Holtz, 2008). Typical of the advice available to sensitivity reviewers, the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) simply state that “the test review process should include empirical analyses and, when appropriate, the use of expert judges to review items and response formats” (Standard 3.6). Similarly, the *International Guidelines for Test Use* (International Testing Commission, 2000) indicate that “competent test users will make all reasonable efforts to ensure that the tests are unbiased and appropriate for the various groups that will be tested” (p.12), but provide no further direction for determining how to undertake such efforts or when a test has achieved an unbiased/appropriate state.

As exemplified by the backlash one major testing agency received for including a question about reality television on their examination instrument (which test takers perceived as culturally and experientially unfair, Steinberg, 2011), the subjective experience of problematic item content by test reviewers and respondents represents a consequential domain. We present two studies that explore the evaluative nature of the sensitivity review process. In Study 1, signal detection analyses are used to examine the influence of individual difference characteristics on reviewers' accuracy and ability to identify problematic item content. Study 2 directs attention towards test takers and investigates the extent to which the presence of problematic item content adversely influences test performance and reactions.

STUDY 1

Despite the regularity with which sensitivity reviews are conducted, relatively little empirical work has examined the evaluative cognitive processes that sensitivity reviewers engage in or the extent to which individual differences might influence the quality of their judgments (Engelhard, Hansche, & Rutledge, 1990). To this end, we posit that signal detection theory (SDT) represents a conceptually plausible framework for characterizing this judgment process. SDT is a perception and decision-making model applicable to phenomenon that require individuals to identify the presence of a target characteristic, stimulus, or event (Green & Swets, 1966; Swets, 1973). The model has proven useful in capturing the performance and behavior of individuals across a variety of domains, such as recognition memory (e.g., learned versus new items, Yonelinas & Parks, 2007), jury decision-making (guilty versus innocent defendants, Kerr, 1993), clinical assessment/diagnosis (unwell versus healthy patients, McFall & Treat, 1999), weather forecasting (patterns predictive of bad versus good weather, Mason, 1982), performance appraisal ratings (effective versus non-effective job performance, Lord, 1985), and personnel selection (desirable versus undesirable applicants, Knight & Frederickson, 1982). The primary decision procedure underlying SDT holds that when determining whether a stimulus "signal" is present, individuals combine relevant information about the event into an impression representing the strength of evidence about the presence or absence of that signal. The individual then compares the magnitude of this impression against an internally derived decision criterion. If the perceived evidence exceeds the threshold, the person declares that the target characteristic is present; if it does not exceed this threshold, he or she declares that the target characteristic is absent (cf., Green & Swets, 1966; Harvey, 1992; Macmillan & Creelman, 1991).

In experiments examining SDT, each participant's hit rate (proportion of trials a signal is judged present when it is present) and false alarm rate (proportion of trials a signal is judged present when it is absent) are recorded. These results are used to construct person-specific probability distributions that characterize the likelihood of that individual's ability to distinguish signals from noise. Based on this data, two indicators of the judgment process can be extracted (Swets, 1986): *response tendency* (an individual's overall inclination towards perceiving a signal on any trial) and *accuracy* (an individual's ability to distinguish true signals from true noise).

In the context of the sensitivity review, individuals are asked to read a given item, review it for problematic content, and reach a judgment regarding its appropriateness for inclusion on a test. When examining an item, the reviewer forms an impression of the extent to which it possesses potentially insensitive content and compares this impression against a self-determined threshold reflecting the strength of evidence needed to judge an item problematic. Consequently, problematic item content represents a "signal" stimulus that reviewers try to distinguish from non-problematic content (e.g., Harvey, 1992). SDT thus provides a conceptually reasonable and defensible representation of reviewers' cognitive evaluations during the sensitivity review process. Of further value, the theory provides indices that can be used to assess the accuracy and relative tendencies of individuals, which themselves may be uniquely influenced by various predictors. Below, we posit a number of individual difference variables that may influence the judgment process and, therefore, the quality of a sensitivity reviewer's item evaluations.

Potential Influences on the Sensitivity Review Judgment Process

***Demographics*.** The *minority review strategy* is a commonly advocated technique for selecting individuals to conduct sensitivity reviews (cf., Camilli, 1993; Hood & Parker, 1989; Office for Minority Education, 1980). This approach encourages selecting reviewers from races, sex, and cultural backgrounds that are traditionally underrepresented in the likely population of test takers (e.g., ACT, 2006). The assumption is that members of these groups tend to face more discrimination and insensitivity in their daily experiences, and therefore should be more cognizant of certain biases/unfavorable material than majority individuals (Feldman Barrett & Swim, 1998). By the same token, however, some researchers have argued that members of minority subgroups may also be more likely to feel chronically victimized by discrimination, possibly predisposing them to perceive even innocuous or ambiguous stimuli as problematic (Branscombe, Schmitt, & Harvey, 1999; Foster, 2009). Consequently, although minority members may identify more insensitivity on a test, it is unclear whether this results in more *accurate* reviews or is attributable to minorities employing a less stringent decision criterion when judging whether insensitivity is present in an item (cf., Mael, Connerley, & Morath, 1996).

> *Hypothesis 1: Females will be more accurate at detecting insensitive item content (H1a) and will exhibit greater tendencies to view items as insensitive (H1b) than males. Non-White reviewers will be more accurate at detecting insensitive item content (H1c) and will exhibit greater tendencies to view items as insensitive (H1d) than Whites.*

***Stereotype awareness characteristics*.** An effective sensitivity review requires reviewers to explicitly consider what would be fair and non-offensive to all potential test takers (cf., Ehrlich, 1973). Individuals who are more reactive to and critical of prejudicial/biased stimuli often employ cognitive monitoring/processing strategies that help them overcome socially adopted stereotypes in favor of more equitable perceptions (Devine, 1989). As such, reviewers who possess individual difference characteristics that enhance responsiveness to offensive or stereotypic events, cues, signals, etc. may be more likely to detect the presence of problematic test content. However, this heightened vigilance may also produce higher false alarm rates (i.e., classifying an item that is non-problematic as problematic) during the review process. We consider four such stereotype-related characteristics: gender/ethnic identity, gender/ethnic stigma consciousness, perceived attributions to prejudice, and past experiences with discrimination.

Social identity theory posits that individuals derive a portion of their self-concept from acknowledgement of their affiliation with identifiable social groups (Tajfel & Turner, 1979). To the extent that individuals are more strongly identified with a given group, they are more likely to ascribe greater emotional value and significance to their membership with it (Tajfel, 1981). In the context of sensitivity reviews, individuals with higher *gender* or *ethnic identification* would therefore be expected to recognize and react more strongly to related problematic content than less identified members as a means of preserving their self-concepts (cf., Tajfel & Turner, 1986). Studies show a positive correlation between group identification and perceptions of prejudice among members of devalued groups (e.g., Branscombe et al., 1999; Crosby, Pufall, Snyder, O'Connell, & Whalen, 1989; Dion, 1975; Major, Quinton, & McCoy, 2002), suggesting that more highly identified individuals should be more reactive to offensive test content. However, this reactance may also produce more false alarms as highly identified persons have also been shown to perceive discrimination and prejudice in ambiguous/non-discriminatory situations (Eccelston & Major, 2006; Feldman Barret & Swim, 1998; Major, Quinton, & Schmader, 2003).

Stigma consciousness refers to the extent to which someone expects to be stereotyped by others based on a demographic category (Pinel, 1999). Research on stigma consciousness suggests that individuals who expect to be stereotyped are especially attentive to cues that may threaten their social identity (Major et al., 2002; Miller & Kaiser, 2001; Pinel, 1999, 2004; Steele, Spencer, & Aronson, 2002). For example, Kaiser, Vick, and Major (2006) report that women with higher levels of stigma consciousness directed more attention towards subliminally presented sexist language. Thus, individuals high in gender and/or ethnic stigma consciousness may be especially primed to detect subtly offensive material during the item review process.

When faced with information that could be perceived as discriminatory or offensive to one's self-identity, individuals may attempt to cope with the experience by attributing the information to external prejudicial behavior (e.g., Crocker & Major, 1989; Crocker, Voelkl, Testa, & Major, 1991; Dion, 1975). Over time, a person may internalize this mechanism and develop generalized, trait-like attributional tendencies that are relatively pervasive across situations (e.g., Branscombe, et al. 1999; Williams, Shore, & Grahe, 1998). Feldman Barrett and Swim (1998) argue that these *perceived attributions to prejudice* can influence individuals' evaluative appraisals of situational stimuli by encouraging attributional strategies that shift one's predilections towards a default presumption of discriminatory intentions in uncertain situations. Although a predisposition towards attributing ambiguous behaviors to prejudice may help a sensitivity reviewer be more accurate at detecting subtly problematic content in test items, it may also lead to disproportionately higher false alarm rates.

Lastly, *past experiences with discrimination* may also influence perceptions of items. Evidence supporting the vigilance hypothesis (e.g., Allport, 1954; Feldman Barrett & Swim, 1998) suggests that those who have had more frequent encounters with discrimination tend to be more alert towards prejudice in their daily encounters (Crocker & Major, 1989; Crosby, et al., 1989; Inman & Baron, 1996). Major, Gramzow, et al. (2002) point out that this mechanism is consistent with research that shows repeatedly primed cognitive processes are more easily activated—especially in situations with high uncertainty (Sedikides & Skowronski, 1991). Thus, reviewers who have experienced discrimination may be more likely to employ information monitoring and assessment strategies that better enable them to detect problematic item content; however, they may also be "hyper-vigilant," resulting in skewed response tendencies.

In sum, existing research suggests that reviewers with higher levels of stereotype awareness may be more accurate at detecting problematic item content as it improves one's ability to detect even the subtlest forms of insensitivity. However, heightened responsiveness may also make the internal decision criterion one uses for identifying problematic content less stringent, leading to biased response tendencies that disproportionately classify non-problematic items as insensitive (e.g., Major & Kaiser, 2008).

> *Hypothesis 2: Individuals who are more highly gender/ethnic identified, conscious of gender/ethnic stigmas, more likely to make attributions to prejudice, and have had more experience with discrimination will be more accurate at detecting insensitive item content (H2a) and will exhibit greater tendencies to view items as insensitive (H2b).*

Some research also supports the notion that individuals who belong to minority subgroups often develop belief systems, expectations, and perceptions of prejudice/stereotypes that make them especially cognizant of discriminatory or offensive cues and signals (Allport, 1954; Major & O'Brien, 2005; Major et al., 2002; Miller & Kaiser, 2001; Steele, et al., 2002).

Such research implies that minority members may be more effective sensitivity reviewers primarily because they have developed higher levels of stereotype awareness. This suggests that these characteristics should mediate the relationship between an individual's demographic characteristics and the detection of problematic item content.

> *Hypothesis 3: Gender/ethnic identification, gender/ethnic stigma consciousness, perceived attributions to prejudice, and past experience with discrimination will mediate the relationship between reviewer demographics and review accuracy (H3a) and the relationship between reviewer demographics and response tendencies (H3b).*

Social awareness characteristics. Beyond stereotypes and biases, sensitivity reviewers must also take into account how respondents from different social, cultural, educational, and socioeconomic backgrounds might react to an item's content. Greater understanding, consideration, and respect for others' viewpoints are related to a variety of prosocial outcomes (e.g., Galinksy & Moskowitz, 2000; Richardson, Hammock, Smith, Gardner, & Signo, 1994; Underwood & Moore, 1982). Research suggests that these desirable outcomes are achieved because individuals are compelled to reframe their cognitive and metacognitive processing such that one's own self-value becomes enmeshed with the real or imagined experiences of others, making it easier to "put oneself in another's shoes" (Brislin, Worthley, & MacNab, 2006; Davis, Conklin, Smith, & Luce, 1996; Triandis, 2006). As such, being more socially aware/attuned might also influence the detection of problematic test content, even if that content is not necessarily offensive to the reviewer's own self-concept. We examine two individual difference variables believed to improve (cultural intelligence, empathy) and two variables believed to adversely affect the sensitivity review process (social dominance, status legitimacy beliefs).

Cultural intelligence (CQ) is broadly defined as a person's ability to effectively behave and operate in culturally diverse settings. CQ is a multidimensional construct composed of *metacognitive* (thought processes used to acquire, understand and regulate cultural norms/knowledge), *cognitive* (personally acquired knowledge of norms, practices, and customs of different cultures), and *behavioral* (ability to exhibit appropriate verbal/nonverbal actions towards culturally different people) facets (Earley & Ang, 2003)[2]. Different aspects of CQ share unique relationships with various outcomes, such as drawing appropriate inferences based on cultural values, quality of decisions regarding intercultural interactions, cultural adaptation, and role-prescribed task performance (Ang et al., 2007). Given the theoretical relevance of the CQ facets to sensitivity reviewers' activities, reviewers with higher metacognitive CQ, cognitive CQ, and behavioral CQ should be better equipped to detect problematic test content.

Empathy broadly refers to the manner in which an individual interprets and relates to the past, present, or future experiences of another person (Mead, 1934; Piaget, 1932). Davis (1980) suggested that the experience of empathy is characterized by two distinguishable responses: *perspective taking*, which enables one to intellectually understand another's viewpoint, and *empathic concern*, which enables one to emotionally sympathize with another's circumstances. Empirical investigations have generally supported this dichotomous conceptualization in which individuals' perspective taking and empathic concern facilitate the suppression of negative thoughts towards others, which in turn leads to more prosocial behaviors and cognitions (e.g., Coke, Baston, & McDavis, 1978; Davis et al., 1996; Galinksy & Moskowitz, 2000). Thus, empathy should relate to the accurate detection of problematic test material by influencing one's capacity to intellectually and emotionally relate to the experiences of potential test takers.

> *Hypothesis 4: Individuals with greater cultural intelligence (metacognitive CQ, cognitive CQ, and behavioral CQ) and empathy (perspective taking ability and empathic concern) will be more accurate at detecting insensitive item content.*

While some qualities may improve the effectiveness of test review activities, other traits may adversely influence sensitivity review judgments. *Social dominance orientation* (SDO) is a general attitudinal disposition reflecting one's overall desire for intergroup relations to be hierarchical versus equal (Sidanius & Pratto, 1999). Across a variety of situations, cultures, and political ideologies, SDO has been shown to positively correlate with sexist, racist, and cultural elitist beliefs, and negatively correlate with empathetic, tolerant, and altruistic tendencies (Pratto, Sidanius, Stallworth, & Malle, 1994; Pratto et al., 2000; Sidanius, Devereux, & Pratto, 1992). Similarly, *status legitimacy beliefs* reflect the degree to which an individual endorses ideologies that support culturally accepted status hierarchies (Major et al., 2002; Schmader, Major, Eccleston, & McCoy, 2001; Sidanius & Pratto, 1999). Meta-analytic evidence suggests that status legitimacy beliefs are a significant factor in the extent to which individuals hold favorable perceptions of out-group members (Bettencourt, Dorr, Charlton, & Hume, 2001).

Given that sensitivity reviewers must evaluate whether a test item is fair to a wide variety of test takers who may or may not share their values and beliefs or belong to their social group, SDO and status legitimacy beliefs could negatively influence review accuracy and produce biased response tendencies. Individuals with a stronger SDO and status legitimacy outlook may possess more conservative decision criterion for evaluating item content, thus leading to poorer accuracy at perceiving problematic items as truly problematic and an overall tendency to evaluate most items as non-problematic.

> *Hypothesis 5: Individuals who are more socially dominant and who more strongly endorse status legitimacy beliefs will be less accurate at detecting insensitive item content (H5a) and will exhibit a greater tendency to view items as non-problematic (H5b).*

Lastly, the proposed social awareness characteristics may moderate the relationship between demographics and outcomes of the sensitivity review process. Specifically, if the conceptual rationale behind the minority review strategy is true, reviewers from majority subgroups who are more culturally intelligent and empathetic should produce more accurate reviews than majority members who are not. With respect to SDO and status legitimacy perceptions, research on perceptions of discrimination, domain devaluation, and out-group bias suggests that although majority group members are more likely to be socially dominant and status legitimizing, minority members who exhibit higher SDO-like and status legitimacy beliefs tend to place comparatively less value on the importance of achieving equality with majority members (e.g., Bettencourt et al., 2001; Schmader et al., 2001). As a result, such individuals appear more likely to try to “minimize” perceptions of discrimination and therefore may overlook many actually insensitive displays (cf., Major & Kaiser, 2008). Thus, although the relationship between SDO and status legitimacy and review accuracy/response tendency is expected to follow the simple main effect hypothesized above, this effect may be more prominent for minority reviewers.

Hypothesis 6: Cultural intelligence and empathy will moderate the relationship between demographics and reviewer accuracy such that the relationship between these characteristics and reviewer accuracy (H6a) and response tendency (H6b) will be stronger for majority than minority reviewers. Social dominance orientation and status legitimacy beliefs will moderate the relationship between demographics and review accuracy such that the relationship between these characteristics and reviewer accuracy (H6c) and response tendency (H6d) will be significantly stronger for minority than majority reviewers.

In sum, sensitivity reviews appear to share many similarities with judgment/decision-making processes wherein individuals attempt to evaluate a stimulus on the basis of an underlying decision criterion (Harvey, 1992). Study 1 leverages the analytic advantages afforded by SDT to investigate the manner by which certain individual difference characteristics influence sensitivity reviewers' judgments. We empirically examine predictions regarding individuals' propensity to perceive problematic item content based on demographics (Office for Minority Education, 1980), responsiveness to stereotypes and other forms of discrimination (Devine, 1989), and perceptions of one's self-concept and societal norms (Ehrlich, 1973).

Method

Participants

Undergraduate students (*n* = 329) recruited from psychology courses at a large Midwestern university participated in the study for course credit. The sample was primarily young (*M* = 19.62, *SD* = 1.75), White (84%) females (73%). Given low numbers of participants of races other than White, race was collapsed into a dichotomous White/non-White variable.

Procedure

Individuals enrolled in the study through an online recruitment system. After providing initial consent, participants completed online measures of their demographic, stereotype awareness, and social awareness characteristics. Lastly, participants were scheduled to complete a sensitivity review task in person at a later date.

The test reviews were completed individually during large group sessions of 20-50 participants. Prior to beginning the review task, individuals read a 1-page handout describing the purpose and process of sensitivity reviews. The handout, which participants kept and were encouraged to reference during the review task, also provided seven categories/types of insensitive item content distilled from widely available sensitivity guidelines (e.g., ACT, 2006; ETS, 2009); Table 1 provides definitions and exemplar items from this typology. Participants were told that they would be conducting a sensitivity review for a newly developed test of general intelligence designed to assess overall level of knowledge. They were told that their task would be to read each test item and response options and indicate the extent to which they believed the item was problematic or not. Participants were instructed to provide a brief explanation for why they thought an item was problematic if they judged it insensitive to some degree. A sample review for a problematic item was presented to demonstrate how to complete the rating task. Upon completion of all ratings, participants were debriefed and thanked.

Materials and Measures

Test materials. The test used for the sensitivity review consisted of 108 verbal ability items similar to those found on common standardized tests (e.g., SAT, ACT). Half (54) of the items were purposefully designed to contain problematic content, while the remaining items were designed to be non-problematic. To construct the problematic items, exemplars were gathered from sensitivity reviewer training materials and collectively sorted by the first and second author into the seven taxonomic categories shown in Table 1. To include items from the entire content space, additional items were created for underrepresented domains by adding insensitive material to existing questions taken from standardized tests or creating new items. The non-problematic items were also selected from standardized tests and training materials.

The problematic and non-problematic items were presented on the test in random order. Additionally, alternate test forms that reversed item order across forms were used (analyses revealed no significant differences in accuracy (*t*(289) = .26, *ns*) or response tendencies (*t*(289) = .23, *ns*) across forms). Participants were randomly provided one of the two test forms, and provided sensitivity ratings for each item on a four-point scale (*1—highly insensitive, 2—moderately insensitive, 3—possibly insensitive, 4—not problematic*).

Individual difference measures. Apart from those used to capture demographics, descriptions for all measures are presented in Table 2. Three points of clarification about these measures are of note. First, the gender and ethnic stigma consciousness scales used in this study were adapted from Pinel’s (1999) 10-item Stigma Consciousness Questionnaire (SCQ). The SCQ captures individuals’ sensitivity towards stigmas relative to a single, homogenous subgroup to which they belong; thus, different versions of the SCQ are designed for different groups (e.g., females, Blacks). However, this level of specificity was not needed in the current study as it was predicted that people who simply tend to be more cognizant of social stigmas overall should exhibit different accuracies and response tendencies during the sensitivity review process than those who are less so. Thus, we used 6 of the SCQ’s original 10 items that generalized well into a broader measure of stigma consciousness towards gender and race; the reduced length and specificity likely contributed to the lower observed coefficient alpha values for these scales.

Second, the perceived attributions to prejudice scale was also adapted from Branscombe et al.'s (1999) original measure that specifically focused on African Americans as the focal discriminated group. To form a more generalized measure, the term "minority" was substituted as the referent group for the scale items. This change may also have attenuated the coefficient alpha of this scale relative to previous administrations.

Lastly, status legitimacy beliefs can be measured via agreement with any ideology that upholds prevailing social/status hierarchies; as such, we chose to examine beliefs in individual upward mobility. As Major et al. (2002) describe, individual upward mobility reflects "the belief that the status hierarchy is permeable and that individuals have the capacity to improve their own individual status" (p. 269). To the extent that people believe in individual upward mobility, they should be more likely to believe that issues of discrimination and prejudice are idiosyncratic concerns that can be overcome through perseverance and are therefore relatively unimportant determinants of one's achievement (Major et al., 2002; Schmader et al., 2001).

Results

Data Preparation

To be included in analyses, participants were required to correctly respond to three items included in the online survey intended to screen out careless responding (e.g., "Please answer 'Disagree' for this question") and complete all item review ratings (required for signal detection analyses). Based on these criteria, 37 participants were excluded, leaving a final sample size of 292. The excluded sample contained a slightly higher percentage of men ($t(327) = 2.37$, $p < .05$); consistent with this sex difference, the dropped respondents tended to report lower levels of gender identification ($t(327) = 5.61$, $p < .001$), perspective taking ($t(327) = 2.14$, $p < .05$) , and empathic concern ($t(327) = 3.35$, $p < .01$), as well as higher levels of social dominance ($t(327) = 3.607$, $p < .01$) and cognitive cultural intelligence ($t(327) = 3.14$, $p < .01$). However, mean sensitivity ratings for problematic and non-problematic items did not differ between the two groups and power analyses based on conventional criteria ($\alpha = .05$, $\beta = .80$) indicated sufficient power to detect small to moderate effect sizes ($d = .20 – .30$) across all hypotheses using the reduced sample.

Expert vs. Sample Ratings

Using signal detection data to evaluate diagnostic accuracy is dependent upon the objective validity of the stimulus materials (Swets, 1988a)—that is, if a signal trial does not actually contain the signal (or a non-signal trial does contain the signal), the signal detection analyses will be biased. Although care was taken when crafting/selecting study materials to ensure that problematic and non-problematic items were appropriately constructed, previous research suggests that the judgments of subject matter experts (SMEs) can serve as reasonable proxies for conclusions about signal presence when stimulus materials may be ambiguous (Pleskac et al., 2011, Swets, 1988b). To this end, a sample of professional sensitivity reviewers (mean number of years serving as a reviewer = 10.67, *SD* = 8.34) were recruited to review and provide ratings for the 54 problematic items used in the experiment; additionally, a smaller sample of graduate students familiar with test construction practices evaluated the entire 108-item test to determine if the selected items could be considered “true” instances of problematic and non-problematic material for purposes of the signal detection analyses.

To classify items into problematic and non-problematic item sets, percent agreement in the observed frequency of the item-level ratings provided by the SME samples was evaluated.

Specifically, an item was classified as a true problematic item if 70%+ of SMEs rated it as *highly insensitive*, *moderately insensitive,* or *possibly insensitive*; similarly, an item that 70%+ of SMEs rated as *not problematic* was classified as a true non-problematic item. A number of alternative criteria for item classification were also considered. However, the present classification system was selected as it maintained conceptual consistency with the rating scale anchors/instructions provided to reviewers while also maximizing the number of test items retained for analysis. Based on this classification strategy, 37 of the original 54 problematic items (69%) and 36 of the original 54 non-problematic items (67%) were retained for use in the signal detection analyses.

Table 3 presents means and standard deviations of the final items included in the problematic and non-problematic item sets across the professional (n = 31), graduate student (n = 9), and Study 1 participant (n = 292) samples. Both the graduate student SMEs (t(71) = 18.37, p < .001, d = 4.29) and Study 1 participants (t(71) = 10.98, p < .001, d = 2.60) rated the problematic items as significantly more insensitive than the non-problematic items on average, suggesting that individuals tended to perceive differences in these items in the expected manner. Both SME samples tended to provide lower ratings on the problematic items than did the Study 1 participants (F_{all}(2,329) = 62.44, p < .001), though no difference in the average ratings of the non-problematic items were observed (t(299) = .27, *ns*). Despite mean differences in item ratings, the rank ordering of ratings for both the problematic (Spearman rank-order correlation (ρ): Professional-Graduate Student = .44; Professional-Study 1 Participants = .56; Graduate Student-Study 1 Participants = .80, all ps < .01) and non-problematic (Graduate Student-Study 1 Participants = .51, all ps < .01) items were reasonably consistent across all samples.

In sum, the SME ratings lend support to the classification of the 83 test items retained for analysis as instances of “truly” problematic and non-problematic items. Although these descriptive analyses do not confirm that any *single* test item was correctly classified, the comparison of the SMEs' and Study 1 participants' data indicated that both groups perceived these sets of items as sufficiently distinctive and in the expected direction. Furthermore, the results suggest that both SMEs and Study 1 respondents were proficient at detecting sensitivity differences in item content, but differed somewhat in their level of reactivity (i.e., SME raters gave more extreme ratings) to problematic content.

Signal Detection Analyses

For signal detection tasks in which participants use rating scales to indicate the presence of a stimulus, receiver operating characteristic (ROC) curves are the preferred method for evaluating respondent accuracy (Metz, 1978; Swets, 1973; Swets, Tanner, & Birdsall, 1961). ROC curves are computed by plotting each individual's hit rate as a function of the false alarm rate at each anchor of the rating scale and then fitting a curvilinear function through these points. The area beneath the corresponding curve (A_z) represents the respondent's accuracy at detecting signals during the task (Green & Swets, 1966; Stanislaw & Todorov, 1999). Figure 1 presents five ROC curves and their associated A_z estimates for selected participants. A_z ranges from .5 (signals indistinguishable from noise) to 1 (perfect performance) and can be interpreted as the proportion of times a respondent would correctly identify a signal stimulus if signal and noise stimuli were presented simultaneously (Green & Swets, 1966). Thus, the participant in Figure 1 with A_z = .64 would correctly rate as insensitive 64% of all problematic items encountered.

Preliminary examinations of the response data revealed that participants' ratings were not well distributed across all anchors of the response scale. Consequently, an alternative curve-fitting algorithm (the proper binormal model, Metz & Pan, 1999) was used to construct the ROC curves. Relative to more conventional techniques, this algorithm has been shown to produce better fitting ROC curves with more readily interpretable parameters in situations in which signal rating data exhibits restricted range or unusual response patterns (Pesce & Metz, 2006). Computation of the ROC curves and their corresponding measures of A_z were computed using the ROC-KIT software package (available from http://xray.bsd.uchicago.edu/krl)[3].

A variety of indices can be used to capture response tendencies with signal detection data, though *c* is the most highly recommended (Macmillan & Creelman, 1990; Snodgrass & Corwin, 1988). The *c* index represents the distance between a respondent's internal decision criteria and the point at which that individual shows no preference for either an affirmative (i.e., signal is present) or negative (i.e., signal not present) response (Banks, 1970; Stanislaw & Todorov, 1999). The magnitude of *c* signifies the strength of the response tendency, with larger numbers indicating a stronger inclination towards a particular preference. When $c = 0$ no response bias is present; negative *c* values reflect that the individual requires relatively little information to state that a signal was present while positive values reflect that the individual requires relatively large amounts of information to state that a signal was present. In the present study, negative *c* values signify that a participant exhibits a greater tendency towards rating any given item as problematic while positive *c* values signify favoring most items as non-problematic. *c* was calculated in Microsoft Excel using the procedures and formulas outlined in Stanislaw and Todorov (1999)[4].

Table 4 presents the means, standard deviations, and interrcorrelations of the study variables for the final dataset. For all regression analyses, continuous predictor variables were mean-centered prior to their entry in the final equation and the categorical sex and race variables were dummy coded (0 for females and non-Whites, 1 for males and Whites).

Hypothesis Tests

Hypothesis 1a and 1c predicted that females and non-Whites are more accurate reviewers (respectively), while Hypothesis 1b and 1d predicted that these groups will exhibit a greater tendency to view items as problematic (respectively). To test these predictions, accuracy and response tendencies were separately regressed onto participant sex and race. Both predictors accounted for a small but significant proportion of variance in reviewer accuracy (R^2 = .02, F(2,289) = 3.38, p < .05). However, contrary to the presumptions of the minority review strategy, White respondents (M = .78) were slightly more accurate than non-Whites (M = .74) (b = .04, p < .05) while the coefficient for sex failed to achieve significance. With respect to response tendencies, the regression model again explained a small but significant portion of the variance (R^2 = .02, F(2,289) = 3.20, p < .05). In this case, only the coefficient for sex achieved significance (b = .18, p < .05); as predicted, females (M = .44) were significantly more likely to perceive any given item as insensitive relative to males (M = .62).

Relationships between the stereotype-awareness characteristics and respondents' effectiveness during the sensitivity review process were examined next (Hypotheses 2a and 2b). To maintain theoretical continuity among the predictor variables, two separate regressions were conducted for the accuracy and response tendency dependent variables using either the gender- or ethnic-referenced stereotype-awareness variables. Thus for Model 1, past experiences with discrimination was entered at Step 1, gender identification at Step 2, gender stigma consciousness at Step 3, and perceived attributions to prejudice at Step 4; in Model 2, ethnic identification and ethnic stigma consciousness were substituted into Steps 2 and 3, respectively.

For Hypothesis 2a, no variables in Model 1 reached statistical significance. An examination of Model 2 revealed only ethnic identification (b = -.02, p < .05) as a significant predictor (R^2 = .04, F(4,287) = 2.89, p < .05); contrary to predictions, individuals who were more strongly identified with their ethnicity tended to be less accurate. The results for Hypothesis 2b indicated that only gender identification at Step 2 in Model 1 ($\Delta R^2 = .01$, $\Delta F(1,289) = 4.04$, $p < .05$) was significantly related to response tendency ($b = -.11$, $p < .05$) such that individuals with greater gender identification were more likely to perceive insensitivity in any given item (i.e., lower values of *c*). In Model 2, greater ethnic identification ($\Delta R^2 = .02$, $\Delta F(1,289) = 4.43$, $b = -.08$, $p < .05$) was also associated with lower *c* values; however, collinearity problems with ethnic stigma consciousness suggested that the entry order of Steps 2 and 3 be reversed to examine the precise nature of the relationship. When entered at Step 2, higher ethnic stigma consciousness ($\Delta R^2 = .02$, $\Delta F(1,289) = 4.66$, $b = -.12$, $p < .05$) exhibited a similar relationship with response tendency. Together, these findings indicate that ethnic identification and ethnic stigma consciousness were likely significant yet redundant predictors in the model. In sum, the observed results ran partially contrary to the predictions advanced in Hypothesis 2a, though partial support was obtained for Hypothesis 2b for certain stereotype-related characteristic variables (e.g., gender/ethnic identification, ethnic stigma consciousness).

Hypotheses 3a and 3b addressed whether stereotype-awareness characteristics mediate the relationship between reviewer demographics and accuracy and response tendencies, respectively. The indirect effects of sex and race on these outcomes were evaluated using the methodology described by Preacher and Hayes (2008) for conducting regression with multiple mediators. This procedure employs a bootstrapping technique to estimate confidence intervals for the total indirect effect, the specific indirect effect of each mediator, and pairwise contrasts between the specific indirect effects specified in the model.

Table 5 presents the bootstrapped indirect effect estimates for the mediation analyses. An examination of the values for the total indirect effects reveals that only the relationships between race and review accuracy (f = .016, 95% CI = .004 to .031) and race and response tendency (f = .066, 95% CI = .006 to .170) were mediated by the total set of stereotype-related characteristics. In the former case, an examination of the specific indirect effects revealed that ethnic identification was the only variable to significantly mediate the relationship between race and review accuracy (f = .011, 95% CI = .002 to .024) such that non-Whites tended to be more strongly ethnically identified, which was subsequently related to lower accuracy (Hypothesis 3a). For Hypothesis 3b, the specific indirect effects revealed no intermediary variables that were strong enough to mediate the relationship between race and response tendency in the presence of the other predictor variables. Preacher and Hayes (2008) note that this may occur in instances where multicollinearity is an issue and the indirect effects are not substantially larger than zero. Cohen et al. (2003) suggest that a reasonable solution in such instances is to cautiously interpret variables whose significance test is almost large enough to meet the significance criterion as "whatever is lost by the inflation of the Type I error is likely to be compensated by the reduction of the Type II error and the resolution of the apparent inconsistency" (p. 189). In this case, ethnic stigma consciousness (z = 1.42, two-tailed p = .16) and ethnic identification (z = 1.42, two-tailed p = .16) were the two largest potential mediators from the full set. Although the observed significance levels of these tests caution against drawing strong interpretations regarding Hypothesis 3b, the pattern of results was such that non-Whites appeared to report higher levels of both ethnic identification and stigma consciousness, which predicted a greater tendency to perceive any given item as insensitive.

Hypothesis 4 posited that characteristics of cultural intelligence (metacognitive CQ, cognitive CQ, and behavioral CQ) and empathy (perspective taking, empathic concern) relate positively to reviewer accuracy. Hierarchical regression analyses revealed that none of these variables emerged as significant predictors (F(5,286) = .48, *ns*), thus failing to support the hypothesized relationship. Although not formally predicted, these variables also did not account for a significant proportion of variance in participants' response tendencies (F(5,286) = .52, *ns*).

To test Hypothesis 5a and 5b, social dominance orientation and status legitimacy beliefs were regressed onto respondent accuracy and tendencies separately. Regressing accuracy on both predictors accounted for a small but significant proportion of the variance in the outcome variable (R^2 = .02, F(2,289) = 3.93, p < .05). An examination of the coefficients offered partial support for Hypothesis 5a in that more socially dominant individuals tended to be less accurate (b = -.02, p < .05). However, neither variable accounted for significant variance in response tendencies (F(2,289) = 1.00, *ns*), thus failing to support Hypothesis 5b.

The remaining predictions addressed the interaction between demographics and social awareness characteristics on sensitivity review performance (Hypotheses 6a-6d). Cross product interaction terms were first created between the sex and race variables and each of the positive and negative social awareness variables. Hierarchical regression models were then estimated for each of the accuracy and response tendency outcomes, with the Demographic variable entered at Step 1, main effects for the Social Awareness characteristics at Step 2, and the Demographic x Social Awareness interaction terms at Step 3. The tests for Hypothesis 6a and 6b, pertaining to the moderating effect of the cultural intelligence and empathy variables, revealed that only the relationships between race and accuracy (ΔR^2 = .05, ΔF(5,280) = 3.50, p < .01) and race and response tendency (ΔR^2 = .06, ΔF(5,280) = 2.81, p < .05) were significantly moderated; no significant interactions with respondent sex were found. Neither social dominance nor status legitimacy beliefs significantly moderated the relationship between either race or sex and reviewer accuracy or response tendencies (Hypotheses 6c and 6d).

A closer evaluation of the significant interactions observed for Hypothesis 6a revealed that the relationship between race and reviewer accuracy was significantly moderated by behavioral CQ (b = -.09, p < .01) and cognitive CQ (b = .08, p < .01); in the case of Hypothesis 6b, the relationship between race and response tendencies was also significantly moderated by behavioral CQ (b = -.34, p < .05) as well as perspective taking (b = .46, p < .05). Figures 2A through 2D present graphs of the interactions. To evaluate whether the interactions followed the predicted patterns, simple slopes analyses were conducted for each set of interactions (Preacher, Curran, & Bauer, 2006). Results revealed that behavioral CQ had no influence on the accuracy of White reviewers (Fig 2A: $\text{slope}_{\text{White}}$ = -.01, t(288) = .84, *ns*), though non-Whites with higher levels of behavioral CQ were significantly more accurate than non-Whites with lower levels of behavioral CQ ($\text{slope}_{\text{non-White}}$ = .08, t(288) = 2.94, p < .05). However, Whites with higher levels of cognitive CQ were significantly more accurate than Whites with lower levels of cognitive CQ (Fig 2B: $\text{slope}_{\text{White}}$ = .02, t(288) = 1.94, p = .053) while the reverse was true for non-Whites ($\text{slope}_{\text{non-White}}$ = -.06, t(288) = 2.12, p < .05). With respect to Hypothesis 6b, level of behavioral CQ was not significantly related to response tendencies for Whites (Fig 2C: $\text{slope}_{\text{White}}$ = -.01, t(288) = .13, *ns*), though non-Whites with higher levels of behavioral CQ tended to exhibit a greater bias towards perceiving items as non-problematic than non-Whites with lower levels of behavioral CQ ($\text{slope}_{\text{non-White}}$ = .33, t(288) = 2.41, p < .05). Lastly, perspective taking did not significantly influence response tendencies for Whites (Fig 2D: $\text{slope}_{\text{White}}$ = .03, t(288) = .38, *ns*); for non-Whites, higher levels of perspective taking were associated with less biased response tendencies ($\text{slope}_{\text{non-White}}$ = -.43, t(288) = 2.18, p < .05). In total, the pattern of observed interactions were only minimally consistent with Hypothesis 6a (in the case of cognitive CQ), and were not supportive of Hypothesis 6b.

Discussion

The goal of Study 1 was to better understand the manner by which individual difference characteristics influence sensitivity judgments. Using SDT as an analytic framework, a number of characteristics were found to significantly relate to test reviewers' overall accuracy and response tendencies, though effects were generally small. Whites and respondents exhibiting weaker ethnic identification and social dominance orientation tended to be slightly more accurate reviewers, whereas females and respondents who were more strongly gender identified, ethnically identified, and reactive to ethnic stigmas were more likely to perceive any given item as insensitive. The mediation and moderation analyses suggested slightly more nuanced interpretations of these findings, especially those pertaining to non-Whites' reviews. Lower review accuracy for non-Whites appeared to be partially attributable to their higher levels of ethnic identification; furthermore, the extent to which non-Whites reported possessing knowledge of and the ability to adapt to diverse cultural customs exerted a greater influence on the accuracy of their sensitivity reviews. In accounting for response tendencies, there was tentative evidence to suggest that less restrictive decision criterion for non-Whites' was related to higher ethnic identification and heightened awareness of ethnic stigmas. However, the significant interaction effects with observed with the tested social awareness variables revealed this tendency could potentially be offset in non-Whites who possessed a greater capacity for perspective-taking.

Despite generally small effect sizes, the results from Study 1 reveal possible implications for the selection of sensitivity reviewers. In general, reviewers with qualities that facilitate accuracy should be preferred (cf., Green & Swets, 1966; Harvey, 1992). However, the perceived cost of leaving insensitive items on the test (e.g., Camilli, 1993; Haladyna & Downing, 2004;

Ployhart & Holtz, 2008; Ramsey, 1993) suggests that reviewers who adopt a “better safe than sorry” mentality during the review process (i.e., exhibit a greater tendency to perceive items as insensitive and therefore generate more false alarms) may be preferred over those who are less reactive. The tradeoff, however, is that such reviewers may be detrimental to the efficiency of test development as they are more likely to remove and/or suggest unnecessary revisions to non-problematic items. One possibility for assessing the costs of such a tradeoff and gauging the impact of problematic test content is to examine its effects on the psychometric quality of a test. Although there are many reasons for conducting sensitivity reviews (e.g., McPhail, 2010), a common belief is that more sensitive tests will lead to more psychometrically sound evaluations that minimize subgroup performance differences and adverse impact (Ployhart & Holtz, 2008). Attention is directed to this notion in Study 2.

STUDY 2

Arguably the most important concern for sensitivity reviews is their impact on test taking outcomes. The insensitivity of test material might conceivably influence two domains. First, insensitive item content may interfere with the performance of certain groups of individuals and thus lead to observable between-group performance differences (e.g., Camilli, 1993; Jensen, 1980). Second, insensitivity on a test might negatively influence test takers’ perceptions of the test, which could engender poor evaluations of the organization (Hausknecht et al., 2004). The purpose of Study 2 is to investigate the experience of problematic item content from the perspective of test takers and the extent to which the presence of insensitive items differentially influences the performance and perceptions of different subgroups.

Examinations of group performance differences across test items are relatively common, though explanations for why any particular item may lead to performance discrepancies are not always clear. For example, spurious relations elicited by the content, format, or presentation of certain items may introduce construct-irrelevant variance that engenders differential group performance (e.g., reading comprehension requirements on a mathematical ability test negatively influencing test takers with lower verbal ability, Cohen et al., 2003; Haladyna & Downing, 2004). Another possibility could be that specific item cues, contexts, or situations make group membership salient and bring to mind negative domain stereotypes which undermine the performance capacity of certain respondents (e.g., Grand, Ryan, Schmitt, & Hmurovic, 2011; Steele & Aronson, 1995). Regardless of the causes underlying the performance differences, the end result is *item bias*, or systematic error in item and overall test validity associated with factors irrelevant to the test (Jensen, 1980).

Whereas sensitivity reviews rely on the subjective evaluations of reviewers, determinations of item bias are made statistically. One of the most commonly used analyses of item bias is differential item functioning (DIF), which utilizes item response theory (IRT) methodologies to identify whether individuals of equal ability have the same probability of correctly answering a given item. Thus, sensitivity review and DIF analyses are both employed to identify test questions likely to impede test takers from responding in ways that allow for accurate inferences (Zieky, 2006); however, the processes by which items are identified as problematic by sensitivity review versus DIF analyses are fundamentally different (e.g., judgmental vs. statistical; conducted before vs. after test administration) and may yield different results (Camilli, 1993; Ramsey, 1993). For example, a number of studies report that test reviewers perform no better than chance when asked to identify a priori items that will show statistical bias (e.g., Englehard, Hansche, & Rutledge, 1990; Plake, 1980; Sandoval & Miille, 1980).

Research has yet to examine whether the item evaluation activities undertaken by sensitivity reviewers ultimately contribute to test taker performance (Ployhart & Holtz, 2008). To address this issue, we examine whether items flagged as insensitive during a sensitivity review subsequently lead to DIF on a test. Given the results of Study 1 and previous research suggesting that women may be more reactive to problematic item content (Mael et al., 1996), we focus on performance differences across sex as the primary between-group comparison.

> *Hypothesis 7: Problematic items will exhibit differential item functioning across sex such that the items will be more difficult for women than men even when individuals are equal on overall verbal ability.*

Sensitivity reviews are also conducted to mitigate negative reactions towards the instrument (McPhail, 2010). Based on the results of Study 1, individual difference characteristics predictive of responsiveness to insensitive test content may also influence the strength of one's reactions towards a test with problematic content. Therefore, we hypothesize that test taker reactions towards a test containing problematic content will be related to demographic characteristics as well as the stereotype- and social-awareness characteristics identified as meaningful predictors of responsiveness to insensitivity in Study 1.

> *Hypothesis 8: Test takers' perceptions of fairness, opportunity to perform, and appropriateness of the test material will be related to their sex and race (H8a) and to their stereotype-related (gender/ethnic identification, gender/ethnic stigma consciousness) and social awareness (social dominance orientation) characteristics (H8b).*

Method

Participants

Participants were 336 students from a large Midwestern university who participated in the study in exchange for course credit. The sample was composed primarily of young (*M* = 19.45, *SD* = 1.65) White (73.6%) participants and was approximately equally distributed across sex (*n* = 170 males). Given the relatively low numbers of participants of races other than White, race was collapsed into a dichotomous White/non-White variable.

Procedure

Participants provided responses to demographic and individual difference measures through an online survey and were then scheduled to complete an untimed verbal ability test in a supervised group setting at a later date. Informed consent was obtained at test administration. Prior to the test, participants were instructed to complete a verbal skills and comprehension test and to respond to a post-test questionnaire regarding their perceptions of the test. After completing all measures, individuals were debriefed and thanked.

Materials and Measures

Verbal ability test. The verbal ability measure consisted of 30 multiple-choice items distributed across three item formats (sentence correction, reading comprehension, and sentence completion) commonly encountered on standardized tests; Cronbach’s alpha for the full test was .68. To create the assessment instrument, 9 problematic and 21 non-problematic items were drawn from the item sets constructed for Study 1. The non-problematic items were chosen by selecting items with the highest sensitivity ratings from Study 1 participants; the mean SME rating for these items was 3.85 (*SD* = .04), suggesting that the items contained virtually no problematic content. When selecting the problematic items, preference was given to those items rated as most problematic and which exhibited the greatest sex differences in sensitivity ratings; the mean rating provided by Study 1 participants and the professional Study 1 reviewers for these items was 2.53 (*SD* = .16) and 1.58 (*SD* = .35), respectively.

Test taker perceptions. Measures of participants' test reactions included a four-item measure of *test fairness* adapted from Kravitz, Stone-Romero, and Ryer (1997) (e.g., "This test is fair") as well as two subscales from Bauer et al.'s (2001) Selection Procedural Justice Scale (*chance to perform*, 4 items: "I could really show my skills and abilities through this test," and *propriety of questions*, 3 items: "The content of the test seemed appropriate"). Coefficient alphas for the three measures were .76, .81, and .67, respectively.

Individual difference measures. In addition to participants' demographics, gender identification (α = .65), ethnic identification (α = .84), gender stigma consciousness (α = .60), ethnic stigma consciousness (α = .76), and social dominance orientation (α = .90) were assessed using the same scales administered in Study 1 (see Table 2).

Results

Similar to Study 1, two items were inserted in the online survey materials to screen out careless responders, of which 36 respondents failed to answer correctly. *t*-tests revealed that the careless responding sample contained a higher percentage of men ($t(334) = 3.87$, $p < .001$) and non-Whites ($t(319) = 2.23$, $p < .05$). These respondents reported higher levels of social dominance ($t(330) = 5.61$, $p < .001$) and scored significantly lower on the verbal ability test overall ($t(334) = 5.91$, $p < .001$). Given the relatively low loss in power from removing these individuals, they were excluded from all analyses. Descriptive statistics and interrcorrelations for the final sample (n = 300) are presented in Table 6.

DIF Analyses

All items on the verbal ability test were dichotomously scored as correct or incorrect. Given the sample size and number of test items employed in the study, a multifaceted approach was pursued to investigate the presence of DIF in the problematic items. Specifically, DIF was assessed using both the non-parametric Mantel-Haenszel (MH) chi-square statistic (Mantel & Haenszel, 1959) as well as by testing for differences in the item characteristic curve parameters from the best-fitting logistic-IRT model using Lord's chi-square test (Lord, 1980).

Briefly, the MH method assesses the likelihood that a relationship between item response and group membership exists conditioned upon the total test score (Holland & Thayer, 1988). Under the null hypothesis of no DIF, the MH statistic follows a chi-square distribution with one degree of freedom; thus, an item is flagged for DIF if its MH statistic exceeds the critical chi-square associated with this distribution. Logistic-IRT models estimate item characteristic curves for every test item that predict the probability for respondents of a given ability level (θ) to correctly answer a given item (Baker, 2001). For dichotomously scored items, there are three commonly employed logistic-IRT models. The one-parameter (1PL) model estimates item characteristic curves based on a single parameter depicting item difficulty (*b* parameter); the two-parameter (2PL) model estimates an additional parameter depicting an item's capability to discriminate high from low ability examinees (*a* parameter); and lastly, the three-parameter (3PL) model incorporates a pseudo-guessing parameter that corrects for chance performance (*c* parameter) (Hambleton & Swaminathan, 1985). To facilitate model convergence, the guessing parameter for the 3PL model was fixed to c = (1 / number of response options) for each item. To examine the presence of DIF, Lord's chi-square statistic tests the null hypothesis that a given IRT model's item parameters are equivalent across groups of respondents. Evidence of DIF is observed if the critical chi-square value for the test is exceeded with degrees of freedom equal to the number of estimated parameters in the IRT model (Lord, 1980). All IRT and DIF analyses were conducted in R (R Development Core Team, 2008) using the *ltm* (Rizopoulos, 2006) and *difR* (Magis, Béland, Tuerlinckx, & De Boek, 2010) packages.

Prior to testing for DIF, a three-step procedure was followed to determine the best-fitting IRT model for use in the analyses. First, parameter and participant ability estimates were estimated for each of the three IRT models described above. Second, modified parallel analyses (MPA) were conducted to evaluate whether the verbal ability test met assumptions of unidimensional needed to perform the IRT analysis (Drasgow and Lissak, 1983). The MPA procedure computes a tetrachoric correlation matrix from the observed item response score data and then calculates the size of the second eigenvalue from this matrix (corresponding to the second-largest factor/source of explained variance). The estimated IRT model parameters are then used to generate a truly unidimensional synthetic dataset, and the same procedure described above is used to extract the second eigenvalue from this simulated data. Finally, the relative size of the second eigenvalues from the observed and synthetic data are contrasted using a Monte Carlo procedure (k = 100) which approximates a test distribution for assessing the null hypothesis that the two eigenvalues are equal. Results from the MPA using the 1PL (p = .89), 2PL (p = .73), and 3PL (p = .63) estimates all failed to reject the null hypothesis, suggesting that the verbal ability test was sufficiently unidimensional to support the use of the proposed IRT models.

The final step in determining the appropriate IRT model for use in the DIF analyses involved assessing each model's goodness of fit at the item-level. Following the basic procedure described by Stone and Zhang (2003, p. 332), Yen's (1981) chi-squared test for dichotomously scored items (Q_1) was computed. This procedure involves aggregating participants into a small number of subgroups (g = 10) based on their estimated ability levels (θ), constructing observed and expected score response distributions for each ability subgroup, and comparing these values using a traditional chi-squared test with *df* = (number of subgroups – number of estimated model parameters). Results revealed that the 2PL model (two item misfits) was comparatively better fitting across all test items relative to both the 1PL (five item misfits) and 3PL (seven item misfits) models. Although it would be reasonable to expect the 3PL model to fit the data best, the relatively small number of respondents and high level of test performance likely restricted the amount of guessing observed in low-ability respondents (i.e., those who stand to benefit the most from guessing), thereby minimizing the benefit of the fixed pseudo-guessing parameter. Consequently, the 2PL model was used to conduct the remaining DIF analyses.

Only two test items were flagged as exhibiting DIF across males and females by both the MH and IRT analyses; contrary to predictions, both were non-problematic items on which women tended to do better than men. Although failing to achieve statistical significance, the ETS Delta index computed for the MH analyses (commonly used to quantify the effect size of DIF; Holland & Thayer, 1988) did identify two problematic items as “suspicious” based on conventional standards and which tended to favor males (Holland & Thayer, 1985). The 2PL IRT model also revealed significant DIF in five additional items as well, four of which were again non-problematic items that tended to be easier for women. The remaining item flagged by the IRT analyses was a problematic item that appeared to slightly favor males (a = .28, b = -3.66) over females (a = .30, b = -3.27). Examination of the item characteristic curves for this item revealed a small advantage for males over females at low respondent ability estimates; at average ability levels and above, however, the item was slightly easier for women than men. In sum, the overall pattern of results was not strongly supportive of Hypothesis 7.

Hypotheses 8a and 8b were evaluated by examining the pattern of correlation coefficients between test taker reactions (fairness, chance to perform, propriety of questions) and individual differences (sex, race, gender/ethnic identity, gender/ethnic stigma consciousness, social dominance orientation). Only the relationship between perceptions of fairness and race was significant (see Table 6), such that non-White respondents tended to see the test as less fair than Whites ($t(285) = 2.30$, $p < .05$, $d = .29$). Overall then, neither Hypothesis 8a or 8b were supported.

Discussion

Study 2 examined whether the presence of insensitive item content contributed to DIF or negative test perceptions for subgroups shown to be reactive to item insensitivity. In general, items containing problematic content did not appear to function in consistently different fashions for men versus women. This is noteworthy given that four of the nine problematic items used in the study were perceived as significantly more problematic by Study 1 female reviewers compared to males—a fact which should have increased the likelihood of performance discrepancies if males and females were differentially processing/reacting to insensitivity as test takers. While prior research has also shown that the judgmental processes used to evaluate the appropriateness of a particular item are not necessarily indicative of an item's psychometric quality (e.g., Englehard et al., 1990; Plake, 1980; Sandoval & Miille, 1980), this study is unique in specifically including items judged to be insensitive a priori. Lastly, apart from a single correlation between race and test fairness, respondents' reactions to the test were generally not associated with their individual differences, suggesting that problematic content on the exam did not lead to substantially different perceptions across test takers.

General Discussion

Test developers conduct sensitivity reviews for a number of reasons, including attempts to improve test reliability or validity by removing construct-irrelevant material, as a "defensive strategy" for demonstrating fairness and good intentions if legally challenged, and to minimize any concerns about the organization/test instrument that the presence of offensive test content could elicit (McPhail, 2010). Despite such intentions, the paucity of empirical investigations regarding the nature of and outcomes associated with the sensitivity review process has made it difficult to conclude whether such objectives are achieved. The present studies attempted to address this evidentiary gap by uniquely evaluating the manner in which problematic item content is perceived and experienced from the perspectives of both test reviewers and test takers.

Implications

The results of Study 1 identified a small number of reviewer characteristics influential in test reviewers' evaluation of problematic item content and which could ultimately impact the quality of the sensitivity review. Although such information could provide insight into questions regarding the identification, recruitment, and selection of future sensitivity reviewers, the results of the SDT analyses make clear that consensus is required on what constitutes an "effective review." Reviewer accuracy and the ability to objectively separate truly insensitive items from non-problematic ones should be the single best indicator of a good reviewer. Based on the present results, the commonly advocated minority review strategy may not necessarily be a superior approach to producing higher quality test reviews. Although observed effect sizes were small, individual differences related to stereotype and social awareness were more informative predictors of sensitivity review accuracy than demographic characteristics. Thus, outcomes of the sensitivity review process could potentially be improved by selecting reviewers who do not strongly adhere to in-group identities and are more receptive to norms of equality—regardless of demographics or experience.

Apart from objective accuracy though, reviewers' general response tendencies also hold important implications as they can influence the relative efficiency of the review process. While false alarms (i.e., saying an item is problematic when it is actually not) and false negatives (i.e., saying an item is non-problematic when it actually is) are unavoidable outcomes of human reasoning (Feldman Barret & Swim, 1998), determining which tendency is more "tolerable" in a set of reviewers will be based on the value that assessment practitioners place on the perceived costs and benefits of the review process itself. Favoring higher false alarm rates reflects the spirit of most current sensitivity guidelines and represents the "safest" approach to test development as it ensures that fewer problematic items end up on a test. However, this preference can lengthen test production activities and time-to-market estimates by requiring unnecessary item revisions or further development of item pools. Furthermore, as the results of Study 2 demonstrate, whether problematic test content is removed may hold relatively few negative consequences for certain psychometric properties or test taker reactions. A related concern is that higher false alarm rates could result in sensitivity panels needlessly removing items that are actually of *good* psychometric quality and/or which have no great impact on minority group members. Awareness of false alarm and false negative rates on the part of major test developers could be a helpful means of identifying ways of improving efficiency and lowering costs of test development.

Note that we do not wish to imply that sensitivity reviews are inconsequential to the development of valid, fair, and socially acceptable assessments. To the contrary, we support the view that attempting to identify problematic test content through both sensitivity reviews and item bias analyses offer developers complementary, non-overlapping methods for evaluating the quality and appropriateness of their assessments and therefore may be useful for different purposes (Camilli, 1993; Ramsey, 1993). This research simply makes explicit that balancing the risk of spending time and money on creating items to replace existing useful ones versus combating legal charges and negative public attention from an insensitive test is a delicate act that is best guided by continued research into the psychological nature and experience of problematic content.

To this end, the present results also point towards a number of theoretical implications and directions for future research in the area. The taxonomy presented in Table 1 offers a useful starting place for subsequent evaluations of item insensitivity. For example, more focused examinations of reviewer and test taker reactance to specific types of problematic content may reveal useful insights into characteristics that may contribute to higher versus lower false alarm rates and thus where sensitivity reviewer training efforts might best be focused. Additionally, the application of cognitive decision-making models other than SDT to the review process could prove beneficial and reveal further information regarding how reviewers make decisions about the appropriateness of an item and what influences those determinations. The results of Study 1 also suggest that examining DIF across racial subgroups and/or individuals who vary across relevant individual difference characteristics may be a fruitful extension of the results described in Study 2. Lastly, despite the consistent rank ordering of the sensitivity ratings provided by the SME and novice reviewers observed in our research, future work examining the influence of sensitivity reviewer training and instructions on response accuracy and tendencies would be useful in determining the effectiveness of sensitivity review panels composed of job incumbents or others with little test development experience and further elucidating the role of individual differences in the review process (e.g., Harvey, 1992; Swets, 1988; Yonelinas & Parks, 2007).

Limitations

One notable limitation relevant to the interpretation of the present results concerns the use of students as sensitivity reviewers in Study 1. Practice and familiarity with test review procedures undoubtedly contributes to the development of cognitive schemas and decision-making structures that facilitate more experienced test reviewers' item evaluation process. However, sensitivity reviews are not always conducted by seasoned test reviewers or individuals knowledgeable about testing/assessment standards. For example, oftentimes civil service organizations select a group of diverse incumbents and give them the one-time job of reviewing test content for insensitivity; in many ways, utilizing student reviewers in the present study parallels this process of simply asking job incumbents—who are likely also "non-experts"—to provide sensitivity judgments. Nevertheless, efforts were made to improve the quality of the data analyzed in the present studies by removing careless responders. Furthermore, making determinations of item insensitivity does not appear to require special expertise, as evidenced by the relatively high rank-order correlations among the mean item ratings of SME reviewers and student participants.

With respect to Study 2, the greatest concern with the student sample was that the relatively inconsequential experimental setting differed appreciably from that of a selection context where performance is tied to desired outcomes. In such settings, certain groups of individuals may react more negatively to inappropriate items and differential item functioning on problematic items may manifest more readily. Future studies examining performance differences caused by insensitive content may benefit from efforts to heighten the stakes of the test.

Conclusion

Sensitivity reviews are viewed as important to minimizing discriminatory hiring practices, ensuring “due diligence” in the creation of selection materials, and positioning an organization’s HR practices as fair and balanced (Hood & Parker, 1989; McPhail, 2010). The present studies provide some preliminary empirical insights into important questions surrounding the efficacy of the sensitivity review process and individuals’ experience with problematic and insensitive test content. However, there is need for further research on this and other techniques designed to improve the fairness of assessment procedures (cf., Ployhart & Holtz, 2008), especially given the important role of such evaluations in many consequential decisions (e.g., college admissions, employee selection). Only through continued and systematic investigations of best practices in these areas can we hope to develop better practices.

References

ACT. (2006). *Fairness report for the ACT tests*. Iowa City, IA: ACT, Inc.

AERA, APA & NCME. (1999). *The standards for educational and psychological testing*. Washington, D.C.: American Psychological Association.

Allport, G.W. (1954). *The nature of prejudice*. Cambridge, MA: Addison-Wesley.

Ang, S., van Dyne, L., Koh1, C.K., Ng, Y., Templer, K.J., Tay, C., & Chandrasekar, N.A. (2007). Cultural intelligence: Its measurement and effects on cultural judgment and decision making, cultural adaptation and task performance. *Management and Organization Review, 3*(3), 335–371.

Baker, F.B. (2001). *The basics of item response theory (2nd ed.)*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Banks, W.P. (1970). Signal detection theory and human memory. *Psychological Bulletin, 74*, 81-99.

Bauer, T.N., Truxillo, D.M., Sanchez, R.J., Craig, J.M., Ferrara, P., & Campion, M.A. (2001). Applicant reactions to selection: Development of the Selection Procedural Justice Scale (SPJS). *Personnel Psychology, 54*, 387-414.

Bettencourt, B.A., Dorr, N., Charlton, K., & Hume, D.L. (2001). Status differences and in-group bias: A meta-analytic examination of the effects of status stability, status legitimacy, and group permeability. *Psychological Bulletin, 127*, 520-542.

Branscombe, N.R., Schmitt, M.T., & Harvey, R.D. (1999). Perceiving pervasive discrimination among African-Americans: Implications for group identification and well-being. *Journal of Personality and Social Psychology, 77*, 135-149.

Brislin, R., Worthley, R., & MacNab, B. 2006. Cultural intelligence: Understanding behaviors that serve people’s goals. *Group and Organization Management, 31*, 40–55.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 397-418). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed). Mahwah, NJ: Lawrence Erlbaum.

Coke, J., Batson, C., & McDavis, K. (1978). Empathic mediation of helping: A two-stage model. *Journal of Personality and Social Psychology, 36*, 752-766.

Cook, L. L., & Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement: Issues and Practice, 10*, 37-45.

Crocker, J., & Major, B. (1989). Social stigma and self-esteem: The self-protective properties of stigma. *Psychological Review, 96*, 608-630.

Crocker, J., Voelkl, K., Testa, M., & Major, B. (1991). Social stigma: The affective consequences of attributional ambiguity. *Journal of Personality and Social Psychology, 60*, 218-228.

Crosby, F., Pufall, A., Snyder, R.C., O’Connell, M., & Whalen, P. (1989). The denial of personal disadvantage among you, me, and all the other ostriches. In M. Crawford & M. Gentry (Eds.), *Gender’s thought: Psychological perspectives* (pp. 79-99). New York: Springer-Verlag.

Davis, M.H. (1980). Davis, M. A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology, 10,* 85.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*, 113-126.

Davis, M.H., Conklin, L., Smith, A., & Luce, C. (1996). Effect of perspective taking on the cognitive representation of persons: A merging of self and other. *Journal of Personality and Social Psychology*, *70*, 713-726.

Devine, P.G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56,* 5-18.,

Dion, K.L. (1975). Women's reactions to discrimination from members of the same or opposite sex. *Journal of Research in Personality, 9,* 294-306.

Dorfman, D.D., Berbaum, K.S., & Metz, C.E. (1992). Receiver operating characteristic rating analysis: Generalization to the population of readers and patients with the jackknife method. *Investigative Radiology, 27*, 723-731.

Drasgow, F., & Lissak, R.I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology, 68*, 363-373.

Earley, P.C., & Ang, S. (2003). *Cultural intelligence: Individual interactions across cultures*. Palo Alto, Calif: Stanford University Press.

Eccleston, C.P., & Major, B. (2006). Attributions to discrimination and self-esteem: The role of group identification and appraisals. *Group Processes: Intergroup Relations, 9*, 147-162.

Ehrlich, H. J. (1973). *The social psychology of prejudice*. New York: Wiley.

Engelhard, G., Hansche, L., & Rutledge, K.E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education, 3,* 347-360.

ETS. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Educational Testing Service.

ETS. (2009). *ETS guidelines for fairness review of assessments*. Princeton, NJ: Educational Testing Service.

Feldman Barret, L., & Swim, J.K. (1998). Appraisals of prejudice and discrimination. In J.K. Swim & C. Stangor (Eds.), *Prejudice: The target's perspective* (pp. 11-36). San Diego, CA: Academic Press.

Foster, M. D. (2009). Perceiving pervasive discrimination over time: Implications for coping. *Psychology of Women Quarterly, 33*, 172-182.

Galinksy, A.D., & Moskowitz, G.B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology, 78*, 708-724.

Grand, J.A., Ryan, A.M., Schmitt, N., & Hmurovic, J. (2011). How far does stereotype threat reach? The potential detriment of face validity in cognitive ability testing. *Human Performance, 24*, 1-28.

Green, D.M., & Swets, J.A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Haladyna, T.M. & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*, 17-27.

Harvey, L.O. (1992). The critical operating characteristic and the evaluation of expert judgment. *Organizational Behavior and Human Decision Processes, 53*, 229-251.

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*, 639-683.

Hautus, M. (1995). Corrections for extreme proportions and their biasing effects on estimated values of *d'*. *Behavior Research Methods, Instruments, & Computers, 27*, 46-51.

Hillis, S.L., & Berbaum, K.S. (2005). Monte Carlo validation of the Dorfman-Berbaum-Metz method using normalized pseudovalues and less data-based model simplification. *Academic Radiology*, *12*, 1534-1541

Hillis, S.L., Obuchowski, N.A., Schartz, K.M., & Berbaum, K.S. (2005). A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette methods for receiver operating characteristic (ROC) data. *Statistics in Medicine*, *24*, 1579-1607.

Hood, S. & Parker, L.J. (1989). Minority bias review panels and teacher testing for initial certification: A comparison of two states' efforts. *The Journal of Negro Education, 58*, 511-519.

Holland, P.W., & Thayer, D.T. (1985). *An alternate definition of the ETS delta scale of item difficulty (Research Report RR-85-43)*. Princeton, NJ: Educational Testing Service.

Holland, P.W., & Thayer, S.T. (1988). Differential item performace and the Mantel-Haenszel procedure. In H.Wainer & H. Braun (Eds.) *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Inman, M. L., & Baron, R. S. (1996). *Influence of prototypes on perceptions of prejudice. Journal of Personality and Social Psychology, 70*, 727–739.

International Testing Commission. (2000). *International guidelines for test use*. Stockholm, Sweden: International Testing Commission.

Jensen, A.R. (1980). *Bias in mental testing*. New York, NY: Free Press.

Johnstone, C.J., Thompson, S.J., Bottsford-Miller, N.A. & Thurlow, M.L. (2008). Universal design and multimethod approaches to item review. *Educational Measurement: Issues and Practice, 27*, 25-36.

Kaiser, C.R., Vick, B., & Major, B. (2006). Prejudice expectations moderate preconscious attention to cues that threatening to social identity. *Psychological Science, 17,* 332-338.

Kerr, N. (1993). Stochastic models of juror decision making. In R. Haste (Ed.), *Inside the juror: The psychology of juror decision making*. Cambridge, UK: Cambridge University Press.

Knight, J.M., & Frederickson, W.A. (1982). Decision theory in employee selection. *Central Business Review, 1*, 13-20.

Kravitz, D.A., Stone-Romero, E.F., & Ryer, J.A. (1997). Student evaluations of grade appeal procedures: The importance of procedural justice. *Research in Higher Education*, *38,* 699-726.

Krieger, N. (1990). Racial and gender discrimination: Risk factors for high blood pressure? *Social Science Medicine, 30,* 1273-1281.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, R.G. (1985). Accuracy in behavioral measurement: An alternative definition based on raters' cognitive schema and signal detection theory. *Journal of Applied Psychology, 70,* 66-71.

Luhtanen, R., & Crocker, J. (1992). A collective self-esteem scale: Self-evaluation of one's social identity. *Personality and Social Psychology Bulletin, 18*, 302-318.

Macmillan, N.A. (1993). Signal detection theory as data analysis method and psychological decision model. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 21-57). Hillsdale, NJ: Erlbaum.

Macmillan, N. A., & Creelman, C. D. (1990). Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" indexes. *Psychological Bulletin, 107*, 401-413.

Macmillan, N.A., & Creelman, C.D. (1991). *Detection theory: A user's guide*. Cambridge: Cambridge University Press.

Mael, F. A., Connerley, M., & Morath, R. A. (1996). None of your business: Parameters of biodata invasiveness. *Personnel Psychology, 49*, 613-650.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*, 847-862.

Major, B.N., Gramzow, R.H., McCoy, S.K., Levin, S., Schmader, T., & Sidanius, J. (2002). Perceiving personal discrimination: The role of group status and legitimizing ideology. *Journal of Personality and Social Psychology, 82*, 269-282.

Major, B.N., & Kaiser, C.R. (2008). Perceiving and claiming discrimination. In L.B. Nielsen & R.L. Nelson (Eds.), *Handbook of employment discrimination: Rights and realities* (pp. 285-299). New York: Springer.

Major, B.N., & O'Brien, L.T. (2005). The social psychology of stigmas. *Annual Review of Psychology, 56*, 393-421.

Major, B.N., Quinton, W.J., & McCoy, S.K. (2002). Antecedents and consequences of attributions to discrimination: Theoretical and empirical advances. In M.P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 34, pp. 251–330). San Diego, CA: Academic Press.

Major, B.N., Quinton, W.J., & Schmader, T. (2003). Attributions to discrimination and self-esteem: Impact of group identification and situational ambiguity. *Journal of Experimental Social Psychology, 39*, 220-231.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Mason, I.B. (1982). A model for assessment of weather forecasts. *Australian Meteorological Magazine, 30*, 291-303.

McFall, R.M., & Treat, T.A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50*, 215-241.

McPhail, S.M. (2010). *Rationales for conducting item sensitivity reviews*. Symposium presented at the meeting of the Society for Industrial and Organizational Psychology, Atlanta, GA**.**

Mead, G.H. (1934). *Mind, self, and society*. Chicago: University of Chicago Press.

Metz, C.E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine, 8,* 283-298.

Metz, C.E., & Pan, X. (1999). "Proper" binormal ROC curves: Theory and maximum-likelihood estimation. *Journal of Mathematical Psychology, 43*, 1-33.

Miller, C.T., & Kaiser, C.R. (2001). A theoretical perspective on coping with stigma. *Journal of Social Issues, 57*, 73–92.

Office for Minority Education. (1980). *An approach for identifying and minimizing bias in standardized tests: A set of guidelines.* Princeton, NJ: Educational Testing Service.

Pesce, L.L., & Metz, C.E. (2007). Reliable and computationally efficient maximum-likelihood estimation of "proper" binormal ROC curves. *Academic Radiology, 14,* 814-829.

Piaget, J. (1932) *The moral judgment of the child*. (trans.) London: Kegan Paul, Trench, Trubner.

Pinel, E.C. (1999). Stigma consciousness: The psychological legacy of social stereotypes. *Journal of Personality and Social Psychology, 76,* 114-128.

Pinel, E.C. (2004). You're just saying that because I'm a woman: Stigma consciousness and attributions to discrimination. *Self and Identity, 3*, 39-51.

Plake, B.S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process. *Educational and Psychological Measurement, 40,* 397-404.

Pleskac, T.J., Keeney, J., Merritt, S.M., Schmitt, N., & Oswald, F.L. (2011). A detection model of college withdrawal. *Organizational Behavior and Human Decision Processes, 115,* 85-98.

Ployhart, R. E., & Holtz, B.C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61*, 153-172.

Pratto, F., Liu, J.H., Levin, S., Sidanius, J., Shih, M., Bachrach, H., & Hegarty, P. (2000). Social dominance orientation and the legitimization of inequality across cultures. *Journal of Cross-Cultural Psychology, 31,* 369-409.

Pratto, F., Sidanius, J., Stallworth, L.M., & Malle, B.F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology, 67*, 741-763.

Preacher, K.J., Curran, P. J., & Bauer, D.J. (2006). Computational tools for probing interaction effects in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics, 31,* 437-448.

Preacher, K.J., & Hayes, A.F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40,* 879-891.

R Development Core Team (2008). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Ramsey, P.A. (1993). Sensitivity review: The ETS experience as a case study. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 367-388). Hillsdale, NJ: Lawrence Erlbaum Associates.

Reckase, M.D. (1996). Test construction in the 1990s: Recent approaches every psychologist should know. *Psychological Assessment, 8*(4), 354-359.

Richardson, D. R., Hammock, G. S., Smith, S. M, Gardner, W., & Signo, M. (1994). Empathy as a cognitive inhibitor of interpersonal aggression. *Aggressive Behavior, 20*, 275-289.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response analysis. *Journal of Statistical Software, 17*, 1-25.

Sackett, P.R., Schmitt, N., Ellingson, J.E. & Kabin, M.B. (2001). High-stakes testing in employment, credentialing, and higher education: prospects in a post-affirmative action world. *American Psychologist*, *56*, 302-318.

Sandoval, J., & Miille, M. P. W. (1980). Accuracy of judgments of WISC-R items difficulty for minority groups. *Journal of Counseling and Clinical Psychology, 48,* 249-253.

Scheuneman, J.D., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement*, *27*(2), 109-131.

Scheuneman, J. D., & Grima, A. (1997). Characteristics of quantitative word items associated with differential performance for female and Black examinees. *Applied Measurement in Education, 10,* 299-319.

Schmader, T., Major, B.N., Eccleston, C.P., & McCoy, S.K. (2001). Devaluing domains in response to threatening intergroup comparisons: Perceived legitimacy and the status value asymmetry. *Journal of Personality and Social Psychology, 80,* 782-796.

Sedikides, C., & Skowronski, J. J. (1991). The law of cognitive structure activation. *Psychological Inquiry, 2*, 169–184.

Sidanius, J., Devereux, E., & Pratto, F. (1992). A comparison of symbolic racism theory and social dominance theory as explanations for racial policy attitudes. *Journal of Social Psychology, 132*, 377-395

Sidanius, J., Liu, J., Pratto, F, & Shaw, J. (1994). Social dominance orientation, hierarchy-attenuators and hierarchy-enhancers: Social dominance theory and the criminal justice system. *Journal of Applied Social Psychology, 24,* 338-366.

Sidanius, J., & Pratto, F. (1999). *Social dominance: An intergroup theory of social hierarchy and oppression*. Cambridge: Cambridge University Press.

Spencer S.J., Steele C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35,* 4-28.

Snodgrass, J.G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117,* 34-50.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*, 137-149.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*(5), 797-811.

Steele, C.M., Spencer, S., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 34, pp. 277–341). San Diego, CA: Academic Press.

Steinberg, J. (2011, March 16). SAT's reality TV essay stumps some. *The New York Times*. Retrieved from http://www.nytimes.com/2011/03/17/education/17sat.html.

Stone, C.A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement, 40,* 331-352.

Swets, J.A. (1973). The relative operating characteristic in psychology. *Science, 182*, 990-1000.

Swets, J.A. (1986). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin, 99*, 100-117.

Swets, J.A. (1988a). Measuring the accuracy of diagnostic systems. *Science, 240*, 1285-1293.

Swets, J.A. (1988b). *Signal detection and recognition by human observers.* New York: Wiley.

Swets, J.A., Tanner, W.P., & Birdsall, T.G. (1961). Decision processes in perception. *Psychological Review, 68,* 301-340.

Tajfel, H. (1981). *Human groups and social categories: Studies in social psychology*. Cambridge: Cambridge University Press.

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

Tajfel, H., & Turner, J.C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. Austin (Eds.), *Psychology of intergroup relations* (2nd ed., pp. 7-24). Chicago: Nelson-Hall.

Triandis, H. C. 2006. Cultural intelligence in organizations. *Group and Organization Management, 31*, 20–26.

Underwood, B., & Moore, B. (1982). Perspective-taking and altruism. *Psychological Bulletin, 91*, 143-173.

Williams, K.D., Shore, W.J., & Grahe, J.E. (1998). The silent treatment: Perceptions of its behaviors and associated feelings. *Group Processes and Intergroup Relations, 1*, 117-141.

Yen, W.M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

Yonelinas, A.P., & Parks, C.M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin, 133*, 800-832.

Zieky, M. (2006). Fairness reviews in assessment. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 359-376). Mahwah, NJ: Lawrence Erlbaum Associates.

Footnotes

[1] A detailed report presenting the full response data for all 33 questions is available from the first author by request.

[2] Earley and Ang (2003) also include a fourth facet in their conceptualization labeled *motivational CQ*, which captures the capability to direct attentional resources toward learning about and functioning in intercultural situations. Given its theoretical treatment and previous research showing that the facet is more strongly related with cultural adaptation and task performance in cross-cultural settings (factors not particularly relevant to the sensitivity review process, Ang et al., 2007), this measure was not included in the present analyses.

[3] Problems arise in the computation of ROC curves, A_z, and response tendency metrics in signal detection analyses for cases in which the false alarm rate (F) = 0 or 1 when the hit rate (H) is $0 < H < 1$ (or, similarly, $H = 0$ or 1 if $0 < F < 1$, Stanislaw & Todorov, 1999). In the present study, 11 participants generated a $F = 0$ with an accompanying H between 0 and 1, thus requiring certain corrections be made to the data. To reconcile the issue for computing response tendency metrics, the loglinear approach recommended by Hautus (1995) was employed; this correction requires adding 0.5 to both the number of hits and the number of false alarms and adding 1 to both the number of signal and noise trials before calculating the hit and false alarm rates for all participants, which effectively constrains F and H to an interval between .01 and .99. As previously noted, A_z is the preferred measure of accuracy in signal detection analyses (Swets, 1988a); however, A_z cannot be calculated for data in which there are no false alarms. In such cases, an alternative estimation of accuracy ($A_{d'}$) can be computed that is exactly equal to A_z when the signal and noise distributions are normally distributed and have the same standard deviation (Macmillan, 1993). In the present data, both of these assumptions were met—or very

nearly so—for most every participant in the dataset (results of these analyses are available from the first author upon request); furthermore, the observed correlation between $A_{d'}$ and A_z across all participants in the study was $r = .90$. Consequently, $A_{d'}$ was used as the index for signal accuracy for the 11 participants in which $F = 0$.

[4]When using interval rating data to calculate *c* (or any response tendency index), *k*-1 values may be calculated to represent the response tendency of participants where *k* equals the number of anchors used on the rating scale. Each of these different values reflects a different interpretation for how participants may provide affirmative versus negative responses using the rating scale (see Stanislaw & Todorov, 1999). For example, four values of *c* can be calculated when a five-point rating scale is used: 1) *c* when an affirmative response equals a rating of 1 and a negative response any rating from 2-5; 2) *c* when affirmative equals rating of 1-2, negative 3-5; 3) *c* when affirmative equals 1-3, negative 4-5; and 4) *c* when affirmative equals 1-4, negative equals 5. The decision as to which computed value is interpreted is largely left to the researcher's discretion. Based on the definitions of the scale anchors provided to respondents, the value for *c* reported and used in all analyses in the present study considers an item rating of 1-3 indicative of an affirmative response (i.e., the participant detected insensitivity in the item) and a rating of 4 as indicative of a negative response. This value was chosen as it was most consistent with the experimental instructions in which participants were informed to respond with a rating of 1-3 if they thought the item was problematic and 4 if the item contained no perceivable problems.

Table 1

Typology of insensitivity for tests and example items

Type	No. of items	Example item
Offensive content	7	Which of the following options, if true, would not be a reason for the above stated trend?[a] a. Some whites, believing it's fashionable to be Indian, stretch the truth about their ancestry, claiming 'My grandmother was a Cherokee princess.' b. American society's acceptance and admiration of the Indian heritage has declined. c. When Indians marry non-Indians, their children can rightfully claim Indian ancestry. d. The federal government guarantees members of tribes health care, financial aid for college, hunting and fishing rights, as well as special grants and loans.
Offensive language	9	Some religious officials claim that the ancient Egyptians' history of brutal violence, ritual sacrifices, and worship of non-Christian deities has contributed to the ________ of bloody genocide ravaging eastern Africa.
Emotionally provocative content	11	Which of the following statements, if true, would support the claim above?[a] a. Many single women with children choose not to apply for welfare b. The number of female-headed families is increasing steadily each year c. According to a national survey, single, childless women choose not to have children because they lack monetary resources d. Females who head households and receive welfare payments report that the payments are not adequate to support their children
Portrayal of gender/racial stereotypes	7	Grace Hopper should be an inspiration to female workers everywhere; (A) not only did she prove that a woman could be (B) highly successful in a field dominated by men, and (C) she was able to do so (D) without special treatment or excessive pleas for equality. No error (E).
Unequal referrals to men and women	7	The temperaments of both architects were markedly different; Kevin was reserved and courteous, Joe was ________ and boastful.
Vocabulary unfamiliar to a group	7	In order to _________ a mortgage, an individual should periodically pay his or her lender principal and interest. a. accrue b. amortize c. abscond d. audit e. augment
Content unfamiliar to a group	6	In India, seeing an elephant when one is leaving for a journey is considered ________ because an elephant represents Lord Ganesha, the Indian God who ________obstacles.

[a]Complete item stem not shown.

Table 2

Description of measures used in Study 1

Measure/construct	Source	# of items	*α*	Sample item
Gender identification	Luhtanen & Crocker (1992)	4	.74	Overall, my gender has very little to do with how I feel about myself. (R)
Ethnic identification	Luhtanen & Crocker (1992)	4	.82	My race/ethnicity is an important reflection of who I am.
Gender stigma consciousness	Pinel (1999)	6	.63	Most people have difficulty viewing those who are not of the same gender as equals.
Ethnic stigma consciousness	Pinel (1999)	6	.69	Most people do not judge others on the basis of their race/ethnicity. (R)
Perceived attributions to prejudice	Branscombe et al. (1999)	10	.67	Suppose a minority group member applies for a job for which he/she feels qualified for. After the interview the minority group member learns that he/she didn't get the job.
Past experience with discrimination	Krieger (1990)	6	.92	I have experienced discrimination at school/work because of the social groups I belong to.
Metacognitive cultural intelligence	Ang et al. (2007)	4	.80	I am conscious of the cultural knowledge I apply to cross-cultural interactions.
Cognitive cultural intelligence	Ang et al. (2007)	6	.83	I know the cultural values and religious beliefs of other cultures.
Behavioral cultural intelligence	Ang et al. (2007)	5	.80	I change my verbal behavior (e.g., accent, tone) when a cross-cultural interaction requires it.
Perspective taking	Davis (1980; 1983)	7	.76	I sometimes try to understand my friends better by imagining how things look from their perspective.
Empathic concern	Davis (1980; 1983)	7	.82	I often have tender, concerned feelings for people less fortunate than me.
Social dominance orientation	Sidanius & Pratto (1999)	16	.90	To get ahead in life, it is sometimes necessary to step on other groups.
Status legitimacy	Major et al. (2002)	4	.78	Advancement in American society is possible for all individuals regardless of ethnicity, gender, culture or age.

Note. Responses for all scales were provided on a five-point Likert scale (*1—Strongly disagree* to *5—Strongly Agree*) except for the Perceived attributions to prejudice questionnaire. For this measure, respondents were asked to indicate their belief in the likelihood that an outcome was attributable to prejudice using a five-point scale which ranged from *1—Not at all due to prejudice* to *5—Completely due to prejudice*.

Table 3

Means and standard deviations for professional, graduate students, and Study 1 sample reviewer ratings

	Professional Reviewers			Graduate Students			Study 2 Respondents		
	n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Problematic Item Subset 1 (*k* = 11)	9	2.28_a	.52	9	2.25_a	.56	292	3.09_b	.52
Problematic Item Subset 2 (*k* = 14)	11	1.94_a	.34	9	2.36_a	.58	292	3.03_b	.51
Problematic Item Subset 3 (*k* = 12)	11	2.09_a	.43	9	2.18_a	.46	292	2.98_b	.53
All Problematic Items (*k* = 37)	31	2.09_a	.43	9	2.27_a	.51	292	3.03_b	.48
All Non-Problematic Items (*k* = 36)	--	--	--	9	3.73_a	.19	292	3.76_a	.14

Note. *n* indicates the number of raters who provided unique ratings for each set of items. *k* indicates the number of items in a given item set. For the Professional Reviewers, different groups of raters provided ratings for each item set and no single reviewer provided ratings for all 37 problematic items; for the Graduate Student and Study 2 Sample Reviewers, the same raters provided ratings for all items. Means with different subscripts in the same row are significantly different (p < .05).

Table 4

Means, standard deviations and interrcorrelations for Study 1 variables (n =292)

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Accuracy (A_z)	.77	.11	--																
2. Response Tendency (c)[a]	.48	.53	.03	--															
3. Sex[b]	.25	.43	.05	**.15**	--														
4. Race[c]	.85	.35	**.14**	.01	**-.12**	--													
5. Gender Identification	3.43	.70	-.07	**-.12**	**-.16**	-.02	(.74)												
6. Ethnic Identification	3.03	.81	**-.18**	**-.12**	-.06	**-.23**	**.44**	(.82)											
7. Gender Stigma Consciousness	3.00	.56	-.04	.03	-.06	-.05	**.35**	**.20**	(.63)										
8. Ethnic Stigma Consciousness	2.82	.59	-.11	**-.12**	.05	**-.25**	**.21**	**.32**	**.45**	(.69)									
9. Perceived Attributions to Prejudice	2.55	.41	-.06	-.08	-.05	-.07	**.14**	**.16**	**.15**	**.21**	(.67)								
10. Past Experience with Discrimination	2.38	1.12	-.08	.01	.00	**-.17**	-.01	.05	**.14**	**.22**	-.01	(.92)							
11. Metacognitive Cultural IQ	3.69	.61	-.05	.01	.03	-.10	.11	**.20**	.07	**.14**	.01	.08	(.80)						
12. Cognitive Cultural IQ	2.76	.69	.03	-.01	.00	**-.13**	.05	.09	.05	.01	-.01	.09	**.42**	(.83)					
13. Behavioral Cultural IQ	3.22	.62	.01	.07	-.06	-.06	**.13**	**.18**	**.15**	**.13**	-.02	-.02	**.35**	**.31**	(.80)				
14. Perspective Taking	3.69	.53	.00	-.05	-.10	-.02	-.07	-.03	**-.12**	-.03	.03	.01	**.28**	**.22**	**.17**	(.76)			
15. Empathic Concern	3.88	.56	-.04	.04	**-.33**	.01	-.03	.02	-.03	.01	.09	-.05	**.12**	.09	.08	**.36**	(.82)		
16. Social Dominance	2.12	.55	**-.12**	.04	**.14**	.04	.05	.09	.03	.02	.01	.05	**-.12**	-.08	-.00	**-.28**	**-.36**	(.90)	
17. Status Legitimacy	2.99	.74	-.08	.08	**.13**	.03	-.11	**-.13**	**-.14**	**-.23**	**-.14**	.00	.03	.07	-.08	-.03	-.07	**.12**	(.78)

Note. Numbers in bold represent significant correlations at $p < .05$ or smaller. α coefficients are presented along the diagonal where applicable.

[a] Positive values indicate tendency to state that items are less problematic, negative values indicate tendency to respond that items are more problematic

[b] Dummy coded variable (0 = female, 1 = male).

[c] Dummy coded variable (0 = non-White, 1 = White).

Table 5

Indirect effect estimates from mediation analyses of sex and race on accuracy and response tendency through stereotype-related characteristics (n = 292, Study 1)

	DV: Accuracy			DV: Response Tendency		
	Point estimate	SE	95% CI	Point estimate	SE	95% CI
IV: Sex						
Gender ID	.002	.003	-.003 - .009	.023	.015	.001 - .064
Gender Stig	.000	.001	-.003 - .003	-.008	.009	-.034 - .004
Attributions	.001	.001	-.001 - .006	.005	.009	-.004 - .039
Disc Exp	.000	.001	-.003 - .003	.000	.004	-.006 - .009
Total	.003	.003	-.003 - .011	.019	.017	-.008 - .063
IV: Race						
Ethnic ID	.011	.005	.002 - .024	.033	.022	-.004 - .090
Ethnic Stig	.002	.006	-.011 - .013	.036	.027	-.012 - .098
Attributions	.001	.002	-.001 - .007	.005	.010	-.005 - .043
Disc Exp	.003	.003	-.003 - .011	-.007	.015	-.040 - .021
Total	.016	.007	.004 - .031	.066	.040	.006 - .170

Note. Total represents the combined total indirect effect through all mediators. ID = identification subscale; Stig = stigma consciousness subscale; Attributions = perceived attributions to prejudice subscale; Disc Exp = past experiences with discrimination subscale.

Table 6

Means, standard deviations and interrcorrelations for Study 2 variables (n =300)

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11
1. Verbal test score	21.54	3.83	(.68)										
2. Fairness	3.41	.44	**.15**	(.76)									
3. Chance to perform	3.32	.79	**.25**	**.35**	(.81)								
4. Propriety of questions	3.73	.76	.09	**.51**	**.34**	(.67)							
5. Gender[a]	.47	.50	-.04	.02	-.04	.00	--						
6. Race[b]	.75	.43	.07	**.14**	.02	.06	.03	--					
7. Gender Identification	3.36	.64	-.11	.00	.03	-.04	**-.14**	-.03	(.65)				
8. Ethnic Identification	3.04	.88	**-.17**	-.03	.00	.05	.03	**-.40**	**.42**	(.84)			
9. Gender Stigma Consciousness	2.96	.49	**-.15**	.00	-.06	.00	-.03	-.05	**.44**	**.23**	(.60)		
10. Ethnic Stigma Consciousness	2.84	.58	-.11	-.05	-.03	-.07	-.05	**-.36**	**.23**	**.41**	**.44**	(.76)	
11. Social Dominance	2.07	.58	**-.19**	-.03	-.08	.06	**.27**	.06	.02	**.12**	.07	-.02	(.90)

Note. Numbers in bold represent significant correlations at $p < .05$ or smaller. α coefficients are presented along the diagonal where applicable.

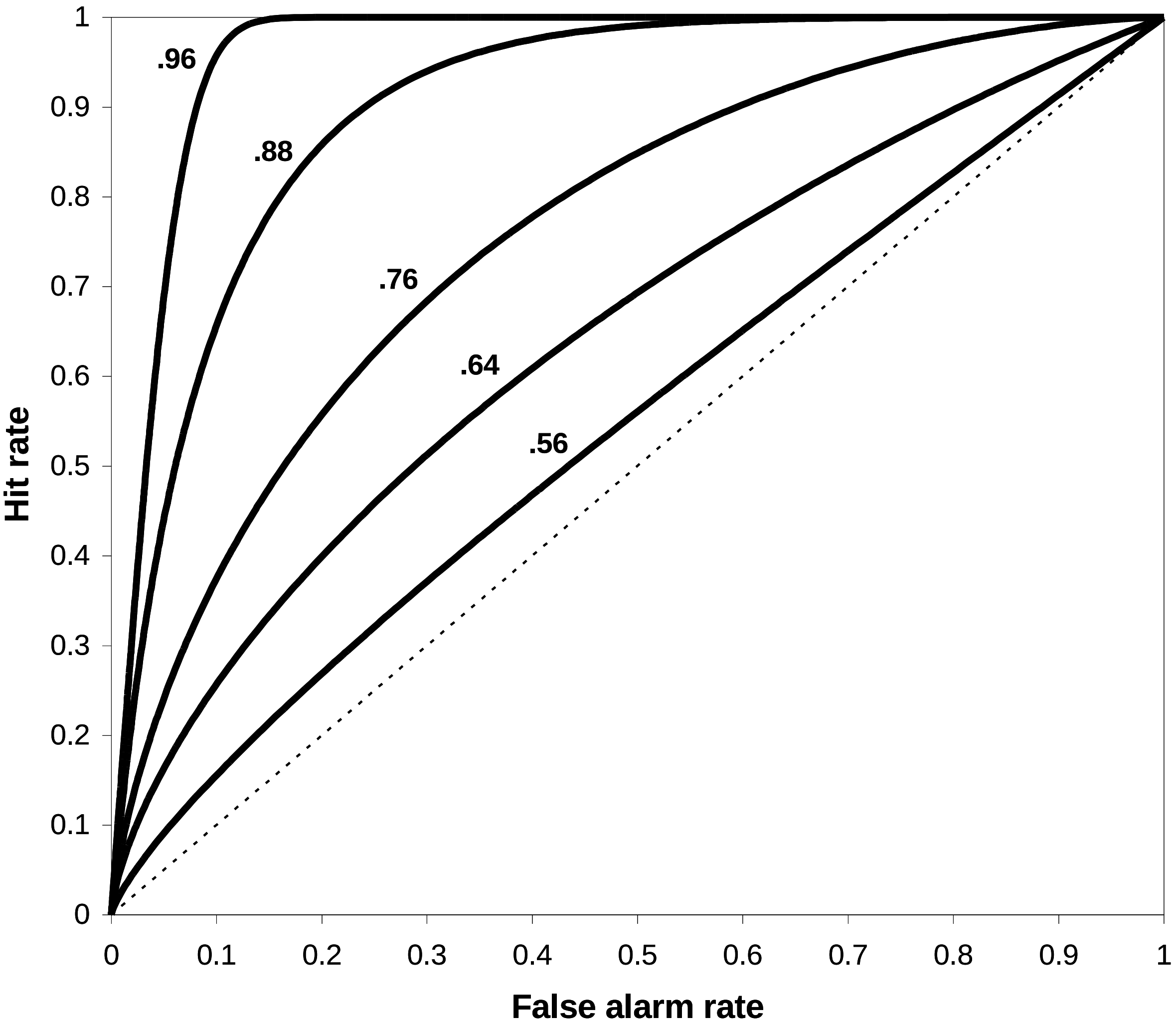
[a] Dummy coded variable (0 = female, 1 = male).

[b] Dummy coded variable (0 = non-White, 1 = White).

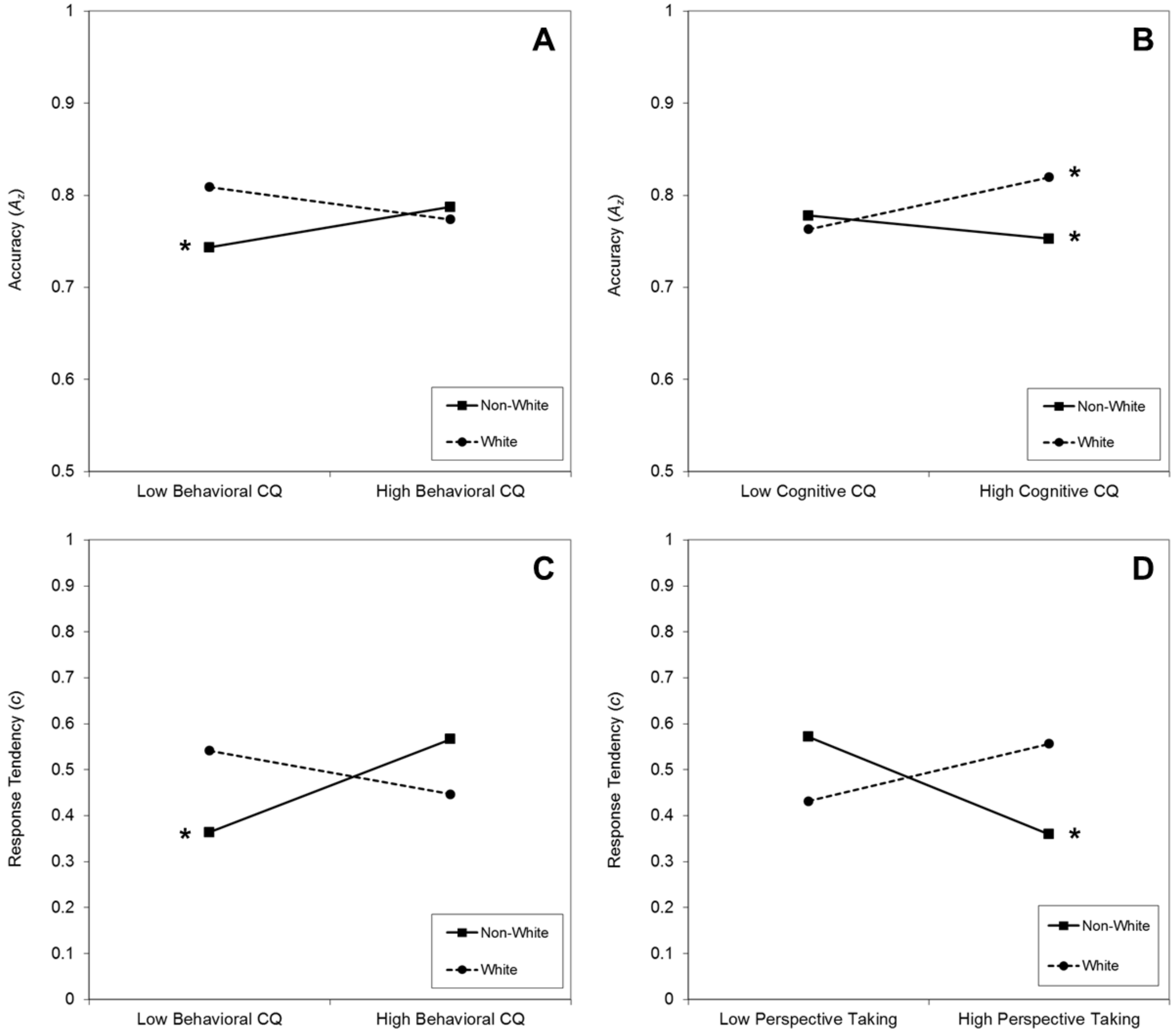
Figure Captions

Figure 1. Example ROC curves for selected Study 1 participants.

Figure 2. Interactions between demographic and individual difference variables on sensitivity review outcomes (Study 1).



Note. Numbers represent the total area beneath each corresponding ROC curve (A_z).



Note. * indicates that the slope of the line is significantly different from zero at p <.05