An Examination of Common Sensitivity Review Practices in Test Development

Juliya Golubovich,
Michigan State University

James A. Grand
The University of Akron

Ann Marie Ryan, and Neal Schmitt
Michigan State University

Citation
Golubovich, J., Grand, J.A., Ryan, A.M., & Schmitt, N. (2014). An examination of common sensitivity review practices in test development. *International Journal of Selection and Assessment, 22*, 1-11.

Abstract

Sensitivity reviews of test content are commonly advocated techniques for reducing bias and enhancing fairness in employment and educational testing. However, few descriptions or empirical investigations of these techniques exist. The present paper presents a study documenting common sensitivity review practices and the extent to which expert reviewers agree in their judgments of item sensitivity. Results indicated that reviewers do not always receive training or adequate guidance and most frequently encounter subtle forms of insensitive item content. Further, only modest agreement in expert ratings of item sensitivity was found. Implications for improving sensitivity review practices are presented.

Keywords: sensitivity review, fairness review, test review

Considering the role that tests play in gaining or denying individuals access to desirable opportunities and institutions, test fairness is a prominent concern for selection specialists. Various methods of improving fairness for different groups of test takers are available (see Ployhart & Holtz, 2008 for a review). Among these is the sensitivity review (also referred to as a bias, or fairness review, ETS, 2009; Ramsey, 1993), which is undertaken to remove any content that could conceivably distract test takers or otherwise prevent them from appropriately demonstrating their true standing on the construct the test is designed to assess (Zieky, 2006). Specifically, sensitivity reviews involve making sure that a test reflects the cultural background of both majority and minority test takers, is accessible format-wise to different subgroups of test takers, and does not contain any inappropriate or offensive content (e.g., sexist, racist, ageist) (ETS, 2002).

Sensitivity reviews are commonly performed as part of large-scale testing programs including, but not limited to, the SAT (College Board, 1998), ACT (ACT, 2008), GMAT (Rudner, 2012), National Assessment of Educational Progress (Ravitch, 2009), California High School Exit Examination (Becker, Wise, Hardoin, & Watters, 2011), Iowa Tests of Basic Skills (Hoover, Dunbar, & Frisbie, 2003), National Council Licensure Examination for Registered Nurses (Wendt, Kenny, & Riley, 2009), California Teacher Licensure Exams (Le & Buddin, 2005), and Singapore Workforce Development Agency's Employability Skills Testing (Jacobsen et al., 2011). Test developers pursue sensitivity reviews with the expectation that the reviews can improve an assessment's psychometric quality, fairness, and legal defensibility (McPhail, 2010). Unfortunately, little is known about typical practices in the area of sensitivity reviews or whether there is agreement regarding what content is considered insensitive. Thus, practitioners seeking

guidance on best practices in conducting sensitivity reviews are often limited to reviewing a few case study descriptions.

To begin to fill this gap, we undertook a survey of professional sensitivity reviewers with the goal of documenting the state of current sensitivity review practices. By professional reviewers, we mean those individuals who engage in these test review activities for major test publishing firms or testing programs on a semi-regular basis, in contrast to those serving in this capacity on a one-time basis for a single exam construction. We investigated reviewers' background and training, the review process, and the problematic content reviewers typically encounter. We also examined the level of expert agreement in flagging items as insensitive. Below, we start by summarizing available information about sensitivity reviews and calling attention to questions that have been raised about their effectiveness. After highlighting the relative lack of available information about common sensitivity review practices, we proceed to describe our survey of professional sensitivity reviewers. Finally, we present survey findings and discuss the implications of these findings for research and practice.

Sensitivity Reviews

Available sources indicate that sensitivity reviews involve one or more testing experts applying "sensitivity" or "fairness" guidelines to a set of test items and making recommendations about what to do with items flagged as problematic (e.g., drop, revise) (Johnstone, Thompson, Bottsford-Miller & Thurlow, 2008; Reckase, 1996). Sensitivity review guidelines (e.g., ACT, 2006; ETS, 2009) provide information about the types of issues for which these experts screen assessments. Examples include content that is stereotypical (e.g., women portrayed only in stereotypic roles, older people represented as senile), offensive (e.g., terms like "crippled" or "fat"), non-inclusive (e.g., terms like "mankind," graphics that lack diversity), provocative (e.g.,

controversial topics like sexuality, abortion, or religion not germane to the purpose of the test), or irrelevant to the test while likely to be differentially familiar to different groups (e.g., idioms or sports terms in a verbal ability test).

There are, however, wide discrepancies with respect to the guidelines that are offered. While some guidelines list specific topics or words that should be avoided (e.g., abortion, suicide, devil, etc.) (Ravitch, 2009; Waters, 2010), others are more general and leave the determination of what may be considered "offensive" up to the individual reviewer (AERA, APA, NCME, 1999). Another discrepancy pertains to whether reviewers should focus primarily on fairness issues in individual items (e.g., representation of gender in particular items, Johnstone et al., 2008), or in the test as a whole (e.g., representation of gender across all items, ACT, 2006). Interestingly, the source of this discrepancy may stem from differences in review procedures. Conceivably, sensitivity reviewers may review entire tests in some contexts and pools of items in others. Ensuring that a group (e.g., women) is represented in as many non-stereotypical (e.g., scientist, lawyer) as stereotypical (e.g., homemaker) roles, as guidelines might suggest (e.g., ETS, 2009), would have limited applicability when reviewing a set of items that may not end up in the same test form or that would comprise a computer adaptive test from which a limited number of items would be presented to any one candidate. One goal of the current study was to gather information about how common reviews of test forms are relative to reviews of item pools, in order to shed additional light on this issue.

While some guidelines describing the problematic content to which reviewers should attend are available, guidelines often are not public but specific to a publisher or testing program. Prescriptive or normative information on reviewer selection, training, and review procedures is even more limited. In the rare cases when such information is shared, it is typically based on a

particular company's internal policies or conventionally endorsed wisdom (cf., ACT, 2006; Camilli, 1993; ETS, 2009; Ramsey, 1993; Waters, 2010). For example, Ramsey (1993) and Becker et al. (2011) describe the sensitivity review process employed by the ETS, including procedures related to the timing of reviews, selection/training of reviewers, and general sensitivity guidelines. Zieky (2006) offers comparable recommendations for conducting systematic sensitivity reviews in licensure testing, such as creating an advisory group specifically for examining fairness issues, providing reviewers with guidelines on how to avoid offensive content and how to use those guidelines, and using a structured/documented approach for completing item reviews (see also Johnstone et al., 2008). It is unclear, however, how representative these few publically available accounts are of sensitivity review practices among test developers and the extent to which their recommendations are followed in various testing domains; one goal of this study is thus to explore the degree to which sensitivity review practices are commonly shared and followed by professional test developers.

Effectiveness of Sensitivity Reviews

The effectiveness of sensitivity reviews at identifying biased or otherwise insensitive items has received some attention from researchers. One method of examining the effectiveness of sensitivity reviews has been to analyze the extent to which reviewers' item evaluations coincide with the results of item bias analyses. A number of studies report that test reviewers perform no better than chance when asked to identify a priori which test items will demonstrate statistical bias (e.g., Broer, Lee, Rizavi, & Powers, 2005; Englehard, Hansche, & Rutledge, 1990; Plake, 1980; Sandoval & Miille, 1980; Young, 2011) or survey items that will be non-equivalent across languages (Carter et al., 2012). Our examination of fifteen books on the subject of assessment suggested that some writers use this evidence as a basis for stating that although

qualitative test reviews are sometimes done, they are not necessarily useful practices, as individuals have not proven effective at identifying biased items. Some have argued that sensitivity reviews and item bias analyses may lead to different test items being flagged as unfair because the two processes are themselves fundamentally different (Camilli, 1993; Ramsey, 1993). Irrespective of whether comparing reviewers' judgments to the results of item bias analyses is an appropriate method of evaluating the effectiveness of sensitivity reviews, there is still no evidence to show that the item evaluation activities of sensitivity reviewers ultimately contribute to test takers' performance (Grand, Golubovich, Ryan, & Schmitt, 2013; Ployhart & Holtz, 2008).

Furthermore, practitioners may be concerned that sensitivity reviews might lead to the removal of good items. Given that cognitive ability is the best predictor of performance outcomes (e.g., Ree & Earles, 1991; Ree, Earles, & Teachout, 1994), if items that happen to have a high cognitive loading get removed during the sensitivity review process in order to be fair to applicants and promote diversity, the validity of the test may suffer (diversity-validity dilemma; Ployhart & Holtz, 2008).

As noted earlier, sensitivity reviews are commissioned for a number of reasons. For example, some test developers would continue to perform these reviews even if they do not necessarily affect test scores and potentially remove valid items simply because being fair to test takers and showing them respect is the socially responsible thing to do (e.g., Ramsey, 1993). However, sensitivity reviews apparently do not always succeed at their goal of ensuring fairness and preventing negative test taker reactions. In a recent case, a major testing agency that is known to perform sensitivity reviews on their test content was publically criticized for

administering an essay question about reality television that many test takers believed was culturally and experientially unfair (Steinberg, 2011).

As one might expect, such cases are generally rare as test developers tend to err on the side of caution when it comes to including test items that could be construed as potentially problematic (e.g., Ramsey, 1993). Nevertheless, the presence of ambiguity and failures in such consequential practices suggest that there may be room for improvement in certain aspects of the review process, such as reviewer selection or process efficiency. As mentioned earlier, some guidelines are vague and leave it up to reviewers to determine what constitutes inappropriate content. The extent of experts' agreement on the implicit definition of item and test sensitivity is not clear, nor is there evidence of inter-rater agreement in judgments of item sensitivity. Guidelines that give reviewers the flexibility to interpret recommendations as they see fit enable idiosyncratic review processes unique to each test development that may function differently and with varying degrees of success. Thus, sensitivity reviewer training and guideline provision might prove to be another area with room for improvement. Unfortunately, lack of information about common sensitivity review practices makes it difficult to address questions that arise about the effectiveness of these practices or to make relevant recommendations for better practices. To this end, we sought to survey professional sensitivity reviewers in an attempt to document the current state of sensitivity review practices in personnel selection and assessment.

Method

Participants

Because there is no one specific background for professional sensitivity reviewers, a multi-pronged approach to participant recruitment was employed by first identifying major test publishing agencies, professional testing associations, and consulting firms that engaged

regularly in test development. For example, we identified SIOP members with expertise in testing and assessment, searched for firms involved in selection and licensure exam development, and gathered membership lists for regional organizations associated with testing. A snowball sampling technique was then used for recruitment by sending e-mails to individuals identified as sensitivity reviewers asking for their participation, and further requesting that these individuals forward the recruitment e-mail to and/or provide contact information for individuals who were known to conduct sensitivity reviews. Note that a number of those initially contacted did not perform sensitivity reviews and thus did not participate in the survey. The final sample consisted of 49 reviewers (55% female, 76% Caucasian), with an average of 10.67 years (*SD* = 8.34) served as a reviewer. About two thirds of the sample estimated that they served as sensitivity reviewers fewer than ten times per year; 16% served as reviewers ten times per year or more. Note that while our sample size may seem small for a survey effort, the pool of individuals who serve in this role on a regular, repeated basis is not large.

Survey Description

The online, anonymous sensitivity reviewer survey contained 33 multiple-choice and free response questions[1]. To develop survey content, we examined available information about sensitivity review practices and identified areas where no or limited information is provided. Survey questions asked about 1) the background and training of reviewers (e.g., preparatory training received, professional background, etc.), 2) the review process (e.g., types of instruments typically reviewed, use of fairness guidelines, primary review activities, etc.), and 3) the nature of insensitive item content encountered (e.g., frequency with which reviewers encounter different types of insensitivity, attention to item-level versus test-level issues, etc.).

[1] Survey items are available from the first author upon request.

Additionally, respondents were asked to provide sensitivity ratings for a subset of 54 problematic test items.[2] We asked respondents to complete this review task in order to examine the extent to which reviewers share a common understanding of what constitutes insensitivity. The following seven categories of insensitivity, derived from fairness guidelines (e.g., ACT, 2006; ETS, 2009), were used to guide the development of the test items: offensive content, offensive language, emotionally provocative content, portrayal of gender/racial stereotypes, unequal referrals to men and women, vocabulary unfamiliar to a group, and content unfamiliar to a group. Items possessing offensive content (7 items) were ones that included information that was unnecessarily upsetting, insulting, or graphic (e.g., drawing an analogy between the effects of an alien species on an ecosystem and of immigrants on American culture, providing estimates of the number of executions carried out in China). Items possessing offensive language (9 items) were ones that included words that could upset or insult test takers (e.g., hell, bloody genocide). Items with emotionally provocative content (11 items) included topics of a sensitive nature that could elicit a negative emotional reaction in test takers (e.g., evolution, slavery, hanging of witches in Salem, Massachusetts). Items containing gender/racial stereotypes (7 items) included content that explicitly or subtly presented stereotypes about certain gender or racial groups (e.g., mention of a girl being unable to handle a hammer, presentation of a successful female as a rare example of her sex). Items possessing unequal referrals to men and women (7 items) had only male or only female subjects (e.g., an item comparing rich men in the Middle East to rich men in America, an item referring to the temperaments of two male architects). Items containing vocabulary unfamiliar to a group (7 items) included terms that are likely to be more accessible to certain groups of test takers (e.g., those of a higher socioeconomic status), such as mortgage-

[2] Ratings for 37 of these items were used in conjunction with data from other sources to support classifying items as problematic or not problematic in Grand, Golubovich, Ryan and Schmitt (2013). Means and SDs of ratings for that subset of items are reported in Table 3 of that paper.

related terms. Finally, items with content unfamiliar to a group (6 items) included information that is likely to be more familiar to certain groups of test takers (e.g., those from a particular cultural background), such as the history of the Great Wall of China and the significance of the elephant as an omen in India.

All the items were verbal ability questions similar to those encountered on common standardized tests (e.g., SAT, ACT, etc.). We created some items by adding insensitive content to items from practice standardized tests or by adapting insensitive items from sensitivity reviewer training materials; other items were developed by the authors for the purposes of this research. To minimize the burden of the review task for reviewers and increase the likelihood of their response to the survey, no single respondent provided ratings for all 54 problematic items; instead, each reviewer provided ratings on one of three non-overlapping item sets containing 18 problematic items. Reviewers provided their ratings on the following four-point scale: *1—highly insensitive, 2—moderately insensitive, 3—possibly insensitive, 4—not problematic.*

Results

Background and training of reviewers

More than half of the respondents indicated that their primary area of expertise was in either organizational psychology (38.8%) or psychometrics (20.4%). The remainder of the sample was distributed across a variety of disciplines, including counseling, education, human resources, linguistics, and business administration, among others. Nearly all individuals felt that they were chosen to participate as a sensitivity reviewer because of their professional capabilities, though a small percentage also indicated that their demographic characteristics likely played a role as well (Table 1). Of note, 71% of the reporting non-White respondents felt that their ethnic background played a role in their recruitment as sensitivity reviewers whereas

only 3% of White respondents believed this was true; 22% of female respondents versus 0% of male respondents felt that their gender contributed to their selection.

Only one third of respondents reported receiving some type of formal training in conducting sensitivity reviews prior to the review process; a slightly greater proportion of individuals considered their education or professional background a source of training (Table 1). Interestingly, 14% of respondents indicated that they had received no training prior to becoming a reviewer and were required to learn the process on their own. When formal training was made available, the most common activities included attending lectures/presentations, discussing sample items, participating in practice review exercises, and covering sensitivity/fairness guidelines in-depth. Respondents who indicated they had taken part in such training efforts reported that the length and specificity of the training sessions varied, with programs lasting anywhere from 30 minutes to upwards of two days. Finally, respondents indicated that activities such as reading relevant literature (e.g., texts, research, manuals; 29%), talking to others (e.g., colleagues, professionals; 14%), and attending meetings/conferences (14%) help them stay abreast of developments relevant to their sensitivity review activities.

Insert Table 1 about here

Description of the review process

Questions pertaining to the actual review process focused on what reviewers were asked to evaluate, how they went about completing their task, and the communication/feedback reviewers experienced. As shown in Table 1, job knowledge and cognitive ability tests were the most commonly reviewed instruments. A wide variety of other assessments were listed with less

frequency including certification exams, integrity tests, assessment center exercises, work samples, and interview questions. The majority (75%) of respondents indicated that they have been asked to review both individual items and entire tests, depending on the situation.

Only half of the respondents (53%) reported receiving a set of sensitivity/fairness guidelines to use when conducting a sensitivity review. However, when asked how they ensure consistency in their reviews, the largest percentage of reviewers (42%) indicated that they relied heavily on guidelines to provide an objective standard. These figures are especially meaningful when one considers that 33% of all respondents reported that their reviews were conducted independently and without any later review or consultation with others. Together, these data underscore the importance of sensitivity guidelines as a relevant informational resource for reviewers. Even when the review task is completed as part of a panel (14% of those sampled) or individually followed by group discussion (24%), well-crafted fairness guidelines likely still provide the most objective procedural aid to ensuring consistency in the process by attenuating the effects of perceptual biases and other social influences that could adversely affect the accuracy of group-based evaluations of an item pool.

A high degree of overlap was observed regarding the nature of the reviews respondents are typically asked to provide (Table 1). The large majority of individuals indicated that their core tasks are to identify potentially inappropriate items and to suggest how they may be improved. When asked about the relative amount of attention they give to item-level versus test-level issues during reviews (Table 1), 40% of respondents indicated giving item-level problems more attention and slightly fewer (38%) indicated trying to be equally attentive to item-level and test-level problems. Only 12% indicated giving test-level problems more attention than item-level problems.

A sizable portion of the sample reported receiving some form of feedback or follow-up communication on their reviews on a regular basis. For example, 53% of reviewers stated that if they recommended an item for revision, they typically saw a revised version of the item at a later time. Similarly, 60% of reviewers reported receiving some form of feedback on their reviews from test publishers (31% "receive feedback;" 29% receive feedback "depending on the situation") while 47% of respondents received information regarding how other reviewers evaluated items in the item pool. Those who received information on other reviewers' item evaluations were split on whether reviewers tend to agree most of the time (50%) or just some of the time (45%). In sum, communication between reviewers and others involved in the process appears to be relatively common.

Nature of insensitive content encountered

As a whole, reviewers reported encountering insensitive item content on a relatively infrequent basis (grand mean = 2.04 across all categories, corresponding to a rating of "somewhat infrequently"), though there was variation in the frequency with which different types of insensitivity were perceived. The pattern of results suggested that the seven types of insensitivity described in Table 1 could be classified into two descriptive categories based on their rates of occurrence (Table 2). The first cluster appears to reflect problematic content that is relatively overt and would be viewed as definitively controversial or derogatory based on widely held societal/cultural standards; subject matter characteristic of this category includes offensive language, offensive content, emotionally provocative content, and stereotypic portrayals of gender/race. The second category consists of more subtly problematic content, including unfamiliar content, unfamiliar vocabulary, and unequal referrals to men and women. Material

characteristic of these items covers topics that typically have less widely shared norms regarding what is socially/culturally "correct" and thus may not be as easily recognized as problematic.

As shown in Table 2, subtly problematic types of insensitivity appear to be more commonly encountered by reviewers on average than overt forms of insensitivity. This difference is not particularly surprising given that the base rates for these types of insensitivity in the item pools which reviewers receive are likely very different. As a number of respondents noted in their open-ended comments, most test writers are diligent enough during the item development stage to avoid including obviously inappropriate terminology and content; however, the ambiguous and subjective nature of subtly problematic content makes it a much more likely candidate to slip through the item writing process undetected. Interestingly, "unequal referrals to men and women" was the least frequently encountered issue among the subtle types of insensitivity. Part of the reason may be that this *test-level* issue has limited applicability in situations where reviewers look at pools of items as opposed to complete tests. Seventeen percent of respondents did indicate that they have only been asked to review pools of items that had not yet been compiled into a test.

Insert Table 2 about here

Biodata questionnaires (14%) and job-related knowledge tests (12%) were cited most frequently as the test instruments with the largest amount of insensitive content. However, caution should be used when interpreting these responses given that respondents varied widely with respect to the types of test instruments they reviewed (e.g., no single respondent indicated having experience reviewing all nine of the tests we asked about) and different instruments may

be subject to different types of insensitivity. Further, a respondent's judgment about a given test instrument's typical level of insensitive content could have potentially been biased by one memorable negative experience with that type of test. Finally, respondents provided some examples of the insensitive item content they had encountered. Examples of offensive language and content included use of words like "retarded" and presenting minorities as the "bad guys". Emotionally provocative topics included abuse and drunk driving. Gender stereotypes included presentation of females in professions such as cook, assistant, and nurse, and of males in leadership roles and professions such as engineer. An example of a racial stereotype encountered in item content was the assumption that minorities commonly require social services. Examples of unfamiliar content were symbols with which immigrants may be less familiar and names of exercise equipment. One respondent also mentioned content that could have different significance in different cultures (e.g., noise tolerance).

Sensitivity ratings

As shown in Table 3, sensitivity reviewers rated a majority of the offensive language items as moderately to highly insensitive, indicating that these items tended to engender the strongest reactions from reviewers. On the other hand, the majority of items featuring unequal referrals to men and women, content unfamiliar to a group, and vocabulary unfamiliar to a group, tended to be seen as not problematic or only possibly insensitive, indicating that that these forms of problematic content generally did not elicit a strong response from reviewers. In fact, none of the unequal referrals and unfamiliar vocabulary items ended up in the moderate to high insensitivity category. Lastly, Table 4 reveals that the sampled reviewers showed only modest agreement and consistency in their ratings of the items they evaluated (ICC coefficients ranged from .31 to .44). These results indicate a relatively sizable amount of between-rater variance in

perceptions of an item's level of sensitivity, suggesting that even experienced professional reviewers may not evaluate a given item's sensitivity/appropriateness in very similar ways.

Insert Tables 3 and 4 about here

As an exploratory analysis, we also examined whether reviewers' demographic or experiential characteristics were related to their average sensitivity ratings on the test items they were asked to evaluate. Consistent with previous research (Grand et al., 2013), professional female reviewers tended to rate items as more problematic than male reviewers (r = -.39, p < .05, d = .85). Sensitivity ratings did not differ as a function of respondents' tenure as a reviewer, ethnicity (minority versus White), professional background (content expert; psychometrician; content expert and psychometrician), frequency of reviews, receiving feedback on reviews, or receiving guidelines. Although they reveal potentially interesting possibilities for subsequent examinations of item reviewers, we do not advocate generalizing strongly from these analyses given that the sample sizes for these exploratory analyses were small (n = 17 to 30) and each reviewer did not rate all possible items.

Discussion

Sensitivity reviews are conducted for a number of reasons, including to remove construct-irrelevant content from the test, ensure fairness for different groups of test takers, and minimize negative test taker reactions (McPhail, 2010). Given the high value placed on these outcomes, sensitivity reviews should be a well-established practice in test development and research. However, the inconsistency in available information and paucity of empirical investigations on sensitivity reviews make it difficult to address questions concerning the

effectiveness of these practices or provide guidelines for best practices. We surveyed sensitivity reviewers to better understand the state of current practice in the hopes of informing research and practice in this area.

This effort revealed a number of interesting findings. First, reviewers demonstrated only modest levels of agreement/consistency in the sensitivity ratings that they provided. This result may partly be reflective of a second key finding: despite their importance as standardizing tools, reviewers do not always receive formal training or guidelines on how to conduct sensitivity reviews. Third, although most individuals believe they are selected as reviewers because of their professional qualifications, ethnic minorities and women were somewhat more likely to attribute their selection to their demographic profiles than Whites and males. Fourth, most reviews appear to be conducted in similar manners, with feedback between/among reviewers and test developers occurring relatively frequently. Finally, although reviewers do not report encountering item insensitivity very frequently, when it is encountered it tends to be more subtle in nature.

Implications

The findings from this study offer a number of implications for practitioners who coordinate sensitivity review practices and for future research in this area. Our results indicate that there is variation in sensitivity reviewer selection and training and that these practices may not always be regulated or systematic. These are areas that can be targeted for improvement given that specific characteristics of reviewers and the manner by which they evaluate insensitivity can potentially influence the quality and efficiency of the sensitivity review process (Grand et al., 2013).

With regard to reviewer selection, our findings suggest that sensitivity reviewers are mostly recruited based on their content expertise, though demographic/cultural background may

also play a role. Ethnic minorities and females were both more likely to attribute their selection as a reviewer to their demographic profiles than Whites and males. This pattern is consistent with minority review strategies that posit selecting minority reviewers (e.g., women, ethnic minorities) helps to ensure that the views and experiences of various test taker groups are represented during the sensitivity review process (Camilli, 1993; Hood & Parker, 1989; Office for Minority Education, 1980). Similar to previous research using student samples (e.g., Grand et al., 2013; Mael, Connerley, & Morath, 1996), we found that professional female reviewers were generally more reactive to insensitive item content. Furthermore, the inter-rater agreement indices presented in Table 4 reveal a substantial degree of variance in item-level sensitivity between raters. Consequently, although the present effort was not intended to specifically investigate such relations, these results and those from existing accounts appear to support the conclusion that the selection and standardization of reviewers has the potential to make a significant impact on the quality of test review practices. This marks an important and practically significant area for future research.

As one point of departure, our results and available sources point to a lack of consideration of individual differences in the selection of sensitivity reviewers. Human decision-making is susceptible to a variety of biases (Kerr, MacCoun, & Kramer, 1996) to which individuals with certain characteristics may be more or less susceptible, and these biases are likely to influence the ratings and decisions made by sensitivity reviewers as well (Ramsey, 1993; Ravitch, 2009; Young, 2011). In Young's (2011) study, for instance, graduate student sensitivity reviewers tended to rate items that they had difficulty answering correctly as more insensitive toward test takers. Attributing one's inability to answer an item correctly to its insensitivity as opposed to one's lack of knowledge may serve as an ego defense mechanism

(Young, 2011). Such biased reviewers, who eliminate items that have a high cognitive loading but which do not contain insensitive content, might reduce the associated test's validity, thereby contributing to test developers' concerns that reviewers might remove valid items from a test.

Relatedly, biased reviewers can hurt the efficiency of the sensitivity review process, which is an important consideration given that test developers sometimes opt to forego conducting a sensitivity review to save time and money (Ramsey, 1993). Importantly, results from our survey indicate that revising items flagged by reviewers as problematic and sending these items back for another review, as well as providing reviewers with feedback on their reviews are not uncommon activities for test developers. Having reviewers flag items that may not be problematic (e.g., items of good psychometric quality, items not likely to lead to negative test taker reactions) for revision or exclusion may be consistent with a "better safe than sorry" review strategy, but this approach can hurt the overall efficiency of the test development process and contribute to longer lead times in selection contexts that may already be pressed for time. Given that there may be individual differences in reviewers' susceptibility to biases, there is value in taking relevant individual differences into consideration when selecting reviewers. To inform practice in this area, future research should investigate reviewers' decision-making process and the individual differences that can influence the quality of the sensitivity review (see Grand et al., 2013 for an example).

When selecting reviewers, practitioners might also want to consider the level of experience individuals have reviewing test content for fairness and sensitivity. A test developer could minimize the costs associated with the sensitivity review process by using more experienced reviewers; experienced reviewers would need less training and calibration than

individuals who have not served as sensitivity reviewers in the past or do not have experience reviewing test content for that particular test developer.

With regard to training, our findings suggest that sensitivity reviewers are not always formally trained prior to reviewing test content, and may instead have to rely on other types of education/training or self-preparatory activities. Providing reviewers with a set of guidelines to follow during the review process should be particularly important when reviewers are not calibrated to a common understanding of fairness and sensitivity via formal training. However, only slightly more than half of the respondents indicated receiving a set of guidelines to follow. The current survey did not investigate the nature of the guidelines respondents received, but existing literature shows that different guidelines provide varying levels of direction on what constitutes unfair or insensitive content. Interestingly, respondents showed only modest levels of agreement and consistency in their ratings of the test items we included in the survey, perhaps reflecting differences in the standards used to evaluate fairness and item sensitivity.

We recommend that sensitivity reviewers always receive formal sensitivity training and a set of guidelines to follow during the review process. Currently, reviewing guidelines, completing practice reviews, and discussing what makes sample items problematic, are training activities that many reviewers do experience—but not all. Such preparation is critical for appropriately preparing and calibrating individuals for the sensitivity review task. Training may, for example, help to minimize the effects of cognitive biases on item reviews, reducing the likelihood that items with a high cognitive loading will be flagged as insensitive just because they are difficult (Young, 2011). We advise that practitioners not rely solely on reviewers' schooling/formal education or experiences as a member of a particular demographic group as adequate preparation for a sensitivity reviewer. Given that conducting training efficiently may be

a practical concern in the context of sensitivity reviews, it would help to focus training strictly on sensitivity as opposed to test content-related issues that would be better addressed by content experts (see Johnstone et al., 2008 for differences between sensitivity and content reviews).

Experienced reviewers should receive refresher training to keep them calibrated for the task and their reviewing skills sharp. New developments in the research literature related to test development and fairness or new legislation can require that existing training materials and guidelines be revised or augmented. Revisions in training materials and guidelines are one of the things that can be brought to reviewers' attention during refresher training. Notably, some survey respondents did not indicate doing anything on their own to keep abreast of new issues and developments, so it would be important for reviewers to receive such information during refresher training.

With regard to guidelines offered to reviewers, we recommend that practitioners explicitly recognize whether the particular review situation calls for reviewers to attend to item-level issues, test-level issues, or both and define their guidelines accordingly. More generally, as available guidelines will sometimes leave the interpretation of fairness up to individual reviewers, perhaps so as to be generic enough to be applicable to different review situations (e.g., different test instruments or content areas), it is important for test developers adopting such guidelines to adapt them to their situation in a way that provides sensitivity reviewers with clearer and more objective standards. Future research could further inform better practices in sensitivity reviewer training by examining the influence of various training and instructional protocols on the effectiveness and efficiency of the review process.

Finally, in considering the modest levels of agreement in item sensitivity ratings between raters in our study, one might ask how many experts a test developer should include in the panel

to achieve a more desirable level of rater reliability. Assuming that additional experts are of the same quality as those already in the sample, the Spearman-Brown correction formula can be applied to show that one would have to increase the number of raters (nine in our example; see Table 4) by a factor of nine or a factor of twenty to raise a reliability of .31 to levels of .80 and .90, respectively. Getting so many experts is likely to be impractical, so we would encourage test developers to enhance rater reliability by focusing instead on calibrating their smaller group of sensitivity reviewers for the review task. Notably, because test developers will often recruit people with diverse experiences and perspectives to serve on a sensitivity review panel, it may not be realistic for them to expect outstanding levels of agreement between reviewers, even if these reviewers have received comprehensive training and clear guidelines.

Limitations

The greatest challenge in this study was identifying individuals who serve as sensitivity reviewers on a repeated basis, which is reflected in the relatively small sample obtained for our survey results. For example, oftentimes civil service organizations may select a group of diverse incumbents and give them the one-time job of reviewing test content for insensitivity or graduate students may occasionally be asked to review test content for sensitivity Because we desired to capture common reviewing practices among professionals who regularly review test content, our findings may have limited generalizability to those ad-hoc panels composed of relatively inexperienced reviewers. We expect, however, that it would be particularly important to train and calibrate those relatively inexperienced reviewers. Finally, variance on the item-rating task might have been restricted by the fact that all the items were created to possess some degree of insensitivity. In spite of its limitations, however, we believe that the current study provides

important information about sensitivity reviews, practices that are fairly common in standardized testing but for few normative or prescriptive accounts are available.

Conclusion

Many view sensitivity reviews as a critical step in the test development process that improves an assessment's psychometric quality, fairness, and legal defensibility, and the associated organization's public image (Hood & Parker, 1989; McPhail, 2010; Ramsey, 1993). The present study provides important information about common practices in sensitivity reviews. By shedding light on common practices, we provide insight into current industry standards and areas where improvements could be implemented. Although more focused empirical research related to sensitivity review practices is beginning to emerge, there is still a need for continued work in this area to evaluate the effectiveness of interventions (e.g., reviewer selection procedures, training protocols) and other similar factors capable of enhancing the rigor, effectiveness, and efficiency of this consequential task in selection and assessment settings.

REFERENCES

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

ACT. (2008). *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.

Becker, D. S., Wise, L. L., Hardoin, M. M., & Watters, C. (2011). Independent Evaluation of the California High School Exit Examination: 2011 Evaluation Report.

Broer, M., Lee, Y. W., Rizavi, S., & Powers, D. (2005). Ensuring the fairness of GRE writing prompts: Assessing differential difficulty. ETS Research Report, RR 05-11. Princeton, NJ: Educational Testing Service.

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.

Carter, N. T., Kotrba, L. M., Diab, D. L., Lin, B. C., Pui, S. Y., Lake, C. J., Gillespie, M. A., Zickar, M. J., & Chao, A. (2012). A comparison of a subjective and statistical method for establishing score comparability in an organizational culture survey. *Journal of Business Psychology, 27,* 451-466.

College Board (1998). *SAT and gender differences*. New York, NY: College Entrance Examination Board and Educational Testing Service

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.

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.

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.

Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2003). *Guide to research and development: Iowa Tests of Basic Skills*. Iowa City, IA: The University of Iowa.

Jacobsen, J., Ackermann, R., Egüez, J., Ganguli, D., Rickard, P., & Taylor, L. (2011). Design of a computer-adaptive test to measure English literacy and numeracy in the Singapore workforce: Considerations, benefits, and implications. *Journal of Applied Testing Technology, 12,* 1-26.

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.

Kerr, N. L., MacCoun, R. J., & Kramer, G. P. (1996). Bias in judgment: Comparing individuals and groups. *Psychological Review, 103,* 687-719.

Le, V-N., & Buddin, R. (2005). Examining the validity evidence for California teacher licensure exams. Santa Monica, CA: RAND Education.

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

McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30-46.

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**.**

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

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.

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.

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.

Ravitch, D. (2009). To be a member of the Governing Board. National Assessment Governing Board.

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

Ree, M. J., & Earles, J. A. (1991). Predicting training success: Not much more than g. *Personnel Psychology, 44,* 321-332.

Ree, M. J., & Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than g. *Journal of Applied Psychology, 79,* 518-524.

Rudner, L. M. (2012). Demystifying the GMAT: Guarding against bias. Graduate Management Admission Council.

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.

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.

Waters, S. (2010). *Practical considerations in developing sensitivity review guidelines.* Symposium presented at the meeting of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Wendt, A., Kenny, L., & Riley, M. (2009). NCLEX fairness and sensitivity review. *Nurse Educator, 34*, 228-231.

Young, C. M. (2011). The influence of person and item characteristics on the detection of item insensitivity. Unpublished doctoral dissertation, University of Akron.

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.

Table 1

Summary of selected survey responses from professional sensitivity reviewers (n = 49)

Characteristics of sensitivity review & reviewers	% of respondents endorsing
Reasons recruited as a sensitivity reviewer	
Professional background	89.8%
Ethnicity	16.3%
Gender	12.2%
Cultural background	8.2%
Other	6.1%
Relevant sensitivity review training received[a]	
Formal schooling/education	35.7%
Formal sensitivity review training	33.3%
No training/self-preparation	14.3%
Psychometric training	9.5%
Item writing training/manual	7.1%
On the job training	4.8%
Professional conferences	2.4%
Manual on bias/sensitivity	2.4%
Legal training	2.4%
Types of tests reviewed	
Job knowledge	63.3%
Cognitive ability	38.8%
Licensing exams	32.7%
Personality	28.6%
Situational judgment tests	28.6%
Educational proficiency exams	20.4%
Biodata	20.4%
Most common sensitivity review activities	
Flag inappropriate items	85.7%
Suggest improvements to problematic items	83.7%
Provide explanations of why item is problematic	42.9%
Edit or rewrite problematic items	12.2%
Rate the insensitivity of items	8.2%
Conduct statistical analyses to identify problematic items	6.1%
Relative attention to item-level versus test-level issues during review[a]	
More attention to item-level issues	40.5%
Equal attention to item-level and test-level issues	38.1%
More attention to test-level issues	11.9%
Not applicable; I do not review full tests	9.5%

Note. Where percentages do not add to 100 in a category, respondents were allowed to select more than one response choice.

[a]$n = 42$

Table 2

Means, standard deviations and correlations for frequency of insensitivity types encountered by professional sensitivity reviewers

Category	Type	*M*	*SD*	1	2	3	4	5	6	7
Overtly problematic	1. Offensive language	1.39	.76	--						
	2. Offensive content	1.50	.66	**.63**	--					
	3. Emotionally provocative content	1.85	.76	**.42**	**.47**	--				
	4. Portrayal of gender/racial stereotypes	2.26	.80	.28	**.42**	**.47**	--			
Subtly problematic	5. Unequal referrals to men and women	2.27	.94	.03	.08	.26	**.52**	--		
	6. Content unfamiliar to a group	2.50	.75	-.09	.25	.18	.22	**.32**	--	
	7. Vocabulary unfamiliar to a group	2.54	.84	-.09	.10	.17	.18	**.39**	**.62**	--

Note. Numbers in bold represent significant correlations at $p < .01$ or smaller. All responses reported on a four-point scale (*1—Never encountered, 2—Encountered somewhat infrequently, 3—Encountered somewhat frequently, 4—Encountered very frequently*)

Table 3
Item ratings by item type

Category	Type	No. of items	Proportion of Items: Moderate to high insensitivity	Proportion of Items: Possible to moderate insensitivity	Proportion of Items: Not problematic to possible insensitivity
Overtly problematic	Offensive language	9	78%	11%	11%
	Offensive content	7	43%	43%	14%
	Emotionally provocative content	11	36%	55%	9%
	Portrayal of gender/racial stereotypes	7	43%	29%	29%
Subtly problematic	Unequal referrals to men and women	7	0%	14%	86%
	Content unfamiliar to a group	6	17%	17%	67%
	Vocabulary unfamiliar to a group	7	0%	43%	57%

Table 4
Means, standard deviations and ICC(2) for sensitivity reviewer ratings

Item Set	*n*	ICC(A,1)	ICC(C,1)	*M*	*SD*
1	9	.31	.36	2.65	.43
2	11	.42	.44	2.25	.30
3	11	.32	.37	2.48	.41
All Items	31	--	--	2.45	.40

Note. *n* reports the number of respondents who provided ratings for each item set. Ratings were provided on a four-point scale (*1—highly insensitive, 2—moderately insensitive, 3—possibly insensitive, 4—not problematic*). Each item set consisted of 18 different problematic items; different groups of raters provided ratings for each item set. ICC(A,1) and ICC(C,1) refer to ICC coefficients reflecting agreement and consistency among raters, respectively (McGraw & Wong, 1996).