Organizational Behavior and Human Decision Processes 121 (2013) 158–173

Contents lists available at SciVerse ScienceDirect

Organizational Behavior and Human Decision Processes

journal homepage: www.elsevier.com/locate/obhdp

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 [a,*], Juliya Golubovich [b], Ann Marie Ryan [b], Neal Schmitt [b]

[a] College of Health Professions and Department of Psychology, The University of Akron, Akron, OH 44325-3701, United States
[b] Department of Psychology, Michigan State University, East Lansing, MI 48824, United States

ARTICLE INFO

Article history:
Received 11 September 2011
Accepted 30 January 2013
Available online 8 March 2013
Accepted by Paul Levy

Portions of this work were presented at the 25th annual meeting of the Society for Industrial and Organizational Psychology, Atlanta, GA and the 26th annual meeting of the Society for Industrial and Organizational Psychology, Chicago, IL

Keywords:
Sensitivity review
Fairness review
Test review
Signal detection analysis
Test bias
Test development
Differential item functioning
Selection
Assessment

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.

Introduction

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

* Corresponding author.
E-mail address: jgrand@uakron.edu (J.A. Grand).

0749-5978/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.obhdp.2013.01.009

criteria that qualifies an item as problematic. For example, sensitivity guidelines often indicate that items with women portrayed in only sex-typed 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 Barret & 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).