Journal of Anxiety Disorders 23 (2009) 256–259

Contents lists available at ScienceDirect

Journal of Anxiety Disorders

Gender bias in the sixteen-item Anxiety Sensitivity Index: An application of polytomous differential item functioning

Nicholas T. Van Dam *, Mitch Earleywine, John P. Forsyth

University at Albany, SUNY United States

ARTICLE INFO

Article history:
Received 30 April 2008
Received in revised form 29 July 2008
Accepted 30 July 2008

Keywords:
Differential item functioning
Anxiety sensitivity
Anxiety

ABSTRACT

Gender differences in measures of anxiety sensitivity (AS) are similar to gender differences across anxiety disorders; females exhibit higher levels of AS and a greater prevalence of anxiety disorders than males. The current study confirms higher scores on the Anxiety Sensitivity Index (ASI) in females. Further analysis reveals, however, that gender differences on the ASI may arise from a single item's bias against women. Four different statistics examining differential item functioning (DIF) indicate that women are more likely to endorse the item, "It scares me when I feel faint", even if they score no higher on the ASI than males. Removing this biased item does not alter internal consistency of the scale, but eliminates the significant gender difference. The results suggest that differences on the ASI require careful interpretation as item bias may artificially inflate ASI scores in females.

1. Introduction

Anxiety sensitivity (AS) is an important cognitive risk factor in anxiety disorders and depression (Vujanovic, Arrindell, Bernstein, Norton, & Zvolensky, 2007). Reiss, Peterson, Gursky, and McNally (1986) initially defined AS as a negative evaluation of anxiety. AS has important implications in the prediction and treatment of clinically significant anxiety episodes (Deacon & Abramowitz, 2006; Vujanovic et al., 2007), especially panic (McNally, 2002). Given the impact of this construct, researchers have emphasized the import of its latent factor structure and validity (e.g., Deacon & Abramowitz, 2006; Zinbarg, Mohlman, & Hong, 1999). AS may also help explain links between anxiety disorders and gender as well.

Gender differences in measures of AS are similar to gender differences across anxiety disorders; females exhibit higher levels of AS and a greater prevalence of anxiety disorders than males (American Psychiatric Association, 2000; Peterson & Reiss, 1993). It is important to establish whether these differences are an artifact of the measurement of these constructs or an actual phenomenon. The Anxiety Sensitivity Index (ASI) has different overall score distributions in males and females. Women score higher on the ASI than men (Peterson & Reiss, 1993) and exhibit a different clustering of responses (Stewart & Baker, 1999; Stewart, Taylor, & Baker, 1997). A twin study reveals that AS is heritable in women but not in men (Jang, Stein, Taylor, & Livesley, 1999). Either AS works differently across the genders, or measures of the construct have the potential for bias. Differential item functioning (DIF), a particular item(s) bias against a group, may underlie the previous findings of gender disparity on the ASI.

DIF is a necessary condition to establish that an item shows bias against one group of participants relative to others. When two groups are equal on their overall pathology, an item that displays DIF is more likely to indicate pathology in one group over the other. The Center for Epidemiologic Studies Depression scale (CES-D; Radloff, 1977) serves as a salient example where an item displays DIF across gender. Women are more likely to endorse an item indicating crying episodes than men who are equally depressed (men with equivalent total scores on the CES-D; Gelin & Zumbo, 2003).

Two types of DIF exist, uniform and non-uniform. Uniform DIF occurs when an item shows equal bias across all levels of a trait for a given group. For example, uniform DIF would occur if men were more likely to endorse a particular item than women regardless of total scale score. That is, men endorse a given item more than women across the whole range of the scale's scores. Non-uniform DIF, on the other hand, occurs when an item only shows bias against one group at a particular trait level or range of scores on the scale measuring that trait. For example, non-uniform DIF would occur if men were more likely to endorse an item when total scale scores were high, but men and women were equally likely to endorse an item if total scale scores were low.

The current study explores responses of men and women on the sixteen-item ASI (Reiss et al., 1986) in a large sample of individuals

* Corresponding author at: University at Albany, Department of Psychology, Social Sciences 369, 1400 Washington Ave, Albany, NY 12222, United States. Tel.: +1 518 445 5533; fax: +1 518 442 4867.
E-mail address: NV122142@albany.edu (N.T. Van Dam).

0887-6185/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.janxdis.2008.07.008

volunteering at the Anxiety Disorders Research Program (ADRP) at the University at Albany, SUNY. The 16-item ASI is the most widely used of the measures of AS (McNally, 2002; Vujanovic et al., 2007). An analysis for potential gender bias on this scale could have meaningful implications for work comparing AS in men and women.

2. Methods

2.1. Procedure

As standard procedure for individuals coming to the ADRP, undergraduate volunteers from a subject pool completed a paper and pencil battery of self-report questionnaires, including the 16-item ASI. The local Institutional Review Board approved this procedure.

2.2. Participants

Prior to completing the assessment battery, candidates were screened for past or present medical or psychological problems via a structured phone interview and subsequently in-person self-report and screening with a modified brief version of the Anxiety Disorders Interview Schedule (ADIS-IV; Brown, DiNardo, & Barlow, 1994). Candidates with serious past or present medical conditions (e.g., cardiovascular issues, asthma, epilepsy, seizures) or psychiatric conditions were excluded. The present research focused on 818 participants who met screening criteria. Half (50.2%) of the participants were male, and participant age ($M = 19.1$, $SD = 1.7$) reflected that the majority of the sample was first-year undergraduates (54.3%), with sophomores (24.4%), juniors (12.9%), and seniors (8.5%) represented in declining proportion. The ethnic/racial distribution was as follows: Caucasian (66%), African American (10.7%), Hispanic/Latino (9.2%), Asian American (7.2%), and other (6.8%).

2.3. Measures

2.3.1. Anxiety sensitivity

Participants completed the 16-item ASI (Reiss et al., 1986) to assess second order anxiety, defined as fear of anxiety-related sensations. Respondents indicated the degree to which individual items characterized them on a 5-point Likert scale ranging from 0 (very little) to 4 (very much). The ASI has good internal consistency ($\alpha = 0.82–0.91$; Peterson and Reiss, 1992) and high test–retest reliability over a 3-year period ($r = 0.71$; Maller & Reiss, 1992).

2.4. Statistical analyses

2.4.1. Methods for detecting DIF

Several statistical approaches can detect DIF in polytomous (non-dichotomous) items, but no specific technique is best in all situations (Mazor, Clauser, & Hambleton, 1992). The principle behind each of the DIF statistics is essentially the same. If DIF is absent, male and female participants with similar anxiety sensitivity should have the same probability of endorsing similar response options on a test item. An item functions differentially when males and females with similar anxiety sensitivity differ significantly in response option endorsement for an individual item. For example, men and women who score a 30 on the total scale should be equally likely to endorse the response '3' on an individual item. Unequal probabilities across groups of response option endorsement suggest DIF. If removing biased items eliminates group differences on the scale, the groups may have differed because of DIF rather than from inherent group differences in the construct.

The most popular classical test theory (CTT) method employed to detect DIF in polytomous items is an extension of the Mantel–Haenszel (MH) statistic, the Mantel chi-square (Mazor et al., 1992). It is based on a 2 (group) $\times$ k (response options) contingency table. For a given item in the ASI, the percentage of males of a certain level of AS selecting a given answer choice (e.g., 3) should be similar to the percentage of females at the same level of AS selecting that same answer choice. The chi-square statistic is based on the difference between observed and expected values in each cell of the table (Penfield, 2007b).

Three other DIF indices are also popular. Two of these employ logic similar to the MH approach. The Liu–Agresti (L–A LOR; Liu & Agresti, 1996) statistic relies on the log-odds ratio of one group selecting a particular response relative to another group, typically yielding a proportion from -1 to 1. The L–A LOR statistic is potentially more robust than other statistics examining DIF and may handle extreme deviations in proportions of responses better than alternative approaches (Penfield, 2007a). Cox's noncentrality parameter estimator (COX's B) parallels the MH approach but relies on the hypergeometric mean (Penfield, 2007b). In the absence of DIF, the odds ratio of responses across gender for each column will be small, suggesting little difference in the proportion of men and women who choose a particular option (see Camilli & Congdon, 1999 for further discussion).

The third popular method for detecting DIF relies on logistic regression. The approach rests on the idea that the score on the item should arise from the true level of the measured construct (which would include any true differences between groups), but not from bias related to group membership. The total score on the questionnaire serves as an indicator of the true level of the measured construct. If a given item is unbiased, the total questionnaire score should serve as the sole significant predictor of the score on the item. Once the total score has predicted the item score, group membership and the interaction of group and total score should no longer account for meaningful variance in the item. The logistic regression approach has an advantage over MH-based statistics for the detection of non-uniform DIF (see Earleywine, 2006; Gelin & Zumbo, 2003, for further discussion).

2.4.2. Current approach

Testing for DIF requires balancing Type I and Type II error rates. Permitting a biased item to remain in a test of psychopathology can have serious consequences, leading some researchers to recommend raising the nominal alpha level as high as 0.20 (Fidalgo, Ferreres, & Muñiz, 2004). This approach improves the power to detect DIF but also increases the chance of flagging an item because of Type I error rather than genuine bias. An alternative way to increase power, while minimizing Type I error, is to use large sample sizes. Recent work reveals that power maximally reaches 0.54 under ideal conditions for samples of 200 or smaller (Fidalgo et al., 2004), creating a danger of missing a biased item because of Type II error. Further simulations suggest that Type II error is as prevalent as 50% for samples of 500 or less, indicating heuristic value of larger samples (Mazor et al., 1992).

The current study employed a multi-step method of initially flagging items for potential DIF using the Mantel chi-square statistic, followed by confirmation of DIF with three other tests (L–A LOR, COX's B, logistic regression). This study also employed a large sample ($N > 800$), and a stringent alpha value of $p < 0.001$. This approach limited Type I error over repeated testing, while conferring enough power to detect any items evidencing DIF. All MH-based statistics were computed using DIFAS 4.0 (Penfield,