Journal of Anxiety Disorders xxx (2008) xxx–xxx

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,

Please cite this article in press as: Van Dam, N.T, et al., Gender bias in the sixteen-item Anxiety Sensitivity Index: An application of polytomous differential item functioning. J Anxiety Disord (2008), doi:10.1016/j.janxdis.2008.07.008

Table 1
Items meeting cut-off criteria for differential functioning across gender

DIF Index	Cut-off statistic	Items (statistic value)
Mantel estimate	$\chi^2 = 10.83$[a]	2 (13.3), 4 (35.7)
Liu–Agresti	|log odds ratio| > 0.64[b]	2 (0.71), 4 (−0.91)
Cox's noncentrality parameter	|log odds ratio| ≥ 0.40[c]	2 (0.40), 4 (−0.54)
Logistic regression	$t = 3.30$[a]	4 (5.88)

Item 2 reads, "When I cannot keep my mind on a task, I worry that I might be going crazy." Item 4 reads, "It scares me when I feel faint."
[a] $p < 0.001$.
[b] Penfield (2007a) large difference criteria.
[c] Approximation of $p < 0.001$ criteria from Camilli and Congdon (1999).

2007b). Logistic regression and all other statistics were computed using SPSS 16.0.

3. Results

3.1. Initial group differences

Men and women did not differ on age, $p > 0.5$. There were small differences in the distribution of ethnicity across gender. A larger percentage of males self-identified as Hispanic or Asian (19.5% of men vs. 13.5% for women); a larger percentage of women self-identified as African American (13.7% of women vs. 7.7% for men), $\chi^2(4, 530) = 11.45$, $p < 0.05$. Despite the obvious heterogeneity of varying minority groups, in the interest of statistical power and study validity, we collapsed all individuals identifying as African American, Asian, Hispanic/Latino, or Other into a single minority group to examine further gender differences by ethnicity. There were no differences between minorities and non-minorities across gender, $\chi^2(1, 803) = 0.10$, $p > 0.9$. The ASI had good internal consistency (Cronbach's $\alpha = 0.89$) and total score varied by gender. Females ($M = 19.2$, $SD = 10.0$) had significantly higher total ASI scores than males ($M = 17.4$, $SD = 9.6$), $F(1, 713) = 5.72$, $p = 0.017$, $d = 0.18$.

3.2. Differential item functioning

The initial Mantel chi-square test of all sixteen items in the ASI revealed DIF at the $p < 0.001$ level for items 2 and 4 (see Table 1). Two MH-based follow-up tests confirmed DIF for both items, though it is notable for item 2 that COX's B was at the cut-value. Item 2 displayed DIF against men, indicating that men are more likely to choose a response one step up the Likert scale relative to women comparable in AS. Statistics for item 4 indicated DIF against women. The logistic regression approach confirmed DIF for item 4 but not item 2. After total ASI score was regressed on item 4, $B = 0.726$, $t(1, 714) = 22.83$, $p < 0.001$, gender still accounted for significant variance, $B = 0.187$, $t(1, 714) = 5.88$, $p < 0.001$, providing evidence for uniform DIF. The interaction was not significant.

3.3. Distributions of potential DIF items

Item 2 shows some evidence of uniform DIF against males. The sexes do not differ on mean response for item 2, $p > 0.1$. This result may stem from low overall item endorsement; skewness was 1.60 with 85.5% of individuals (regardless of gender) choosing a response of 0 (very little) or 1 (a little). Item 4 shows DIF against females. Females ($M = 1.85$, $SD = 1.11$) score significantly higher on this item than males ($M = 1.38$, $SD = 1.12$) $t(1, 765) = 5.89$, $p < 0.001$, $d = 0.42$. The overall range of endorsement for this item was more evenly distributed (skewness of 0.26), with 49.9% of individuals choosing responses of 2 (some): 25.6%, 3 (much): 19.4%, or 4 (very much): 4.9%.

3.4. Impact of item removal on total score

Removing item 2 did not alter the scale's internal consistency (Cronbach's $\alpha = 0.88$), but increased gender differences from $d = 0.18$ to $d = 0.19$ (males, $M = 16.8$, $SD = 9.1$; females, $M = 18.6$, $SD = 9.4$), $F(1, 713) = 6.96$. Removing item 4 did not dramatically alter the scale's internal consistency (Cronbach's $\alpha = 0.88$), but eliminated the significant gender difference on total score, $F(1, 726) = 2.79$, $p = 0.095$; male ASI mean = 16.0 ($SD = 8.9$); female ASI mean = 17.2 ($SD = 9.2$), $d = 0.13$, a 27.7% decrease. Removing both items 2 and 4 did not dramatically alter the scale's internal consistency either (Cronbach's $\alpha = 0.87$), and eliminated the significant difference between genders on overall ASI score, $F(1, 726) = 3.63$, $p = 0.57$; male ASI mean = 15.4 ($SD = 8.4$), female ASI mean = 16.6 ($SD = 8.7$), $d = 0.14$, a 22.2% decrease.

4. Discussion

Females scored significantly higher than males on the ASI, confirming previous work (Peterson & Reiss, 1993; Stewart et al., 1997). DIF analyses revealed a pattern of uniform bias against males on item 2, "When I cannot keep my mind on a task, I worry that I might be going crazy," for 3 out of 4 statistics. Nevertheless, the sexes did not differ on this item. DIF analyses revealed a pattern of uniform bias against females on item 4, "It scares me when I feel faint." Uniform bias against females indicates that regardless of total scale score, females were more likely than males to endorse responses high on the 5-point Likert scale (e.g., 3 or 4) for item 4. Four different methods of analyzing DIF converged upon the same bias at a stringent alpha cut-off, indicating that the results are unlikely to be spurious. Removal of this single item from the total ASI scale had minimal impact on internal consistency but eliminated the difference between genders on the total score, decreasing the size of the effect by more than 25%. These results suggest that any differences in ASI across gender could be due to a bias in the question rather than genuine differences in AS.

Restricted range of responses on item 2, which shows bias against men, may weaken the argument for its removal based on DIF. The strong skew and low endorsement may warrant careful consideration of the item for other reasons. Lack of change in overall ASI score after dropping this item makes it a questionable candidate for an item displaying meaningful DIF. Despite these concerns about the distribution of item 2 responses, when removed alongside item 4, the gender difference on total score is no longer significant and the size of the effect drops 22%. Clearly, further work on gender differences in the ASI can help reveal the extent of the potential for bias in this item.

Content of the biased items suggests that men and women may interpret the items differently. Socialized masculinity stereotypically relies on success, power and competition, restriction of affect, restriction of emotional expression between men, and dedication to work over family; all of these characteristics lead to the potential for a gender role conflict (O'Neil, Helms, Gable, David, & Wrightsman, 1986). Gender role, an internal struggle between social role expectations and personal identity, can lead to higher incidence of psychiatric symptomatology among males (Schaub & Williams, 2007). Endorsing fear is antithetical to the socialized notion of the powerful, emotionally regulated, masculine male. Socialized gender roles may account for differences in item endorsement on the ASI. The item on the ASI that asks about fears of fainting operates differentially; women are much more likely to endorse the item than men are, even when they are equal in AS (see Table 1). This fear is physiologically unwarranted; syncope (fainting) does not vary with gender (Savage, Corwin, McGee, Kannel, & Wolf, 1985). Item 2 also shows bias, though its statistical

Please cite this article in press as: Van Dam, N.T, et al., Gender bias in the sixteen-item Anxiety Sensitivity Index: An application of polytomous differential item functioning. J Anxiety Disord (2008), doi:10.1016/j.janxdis.2008.07.008

basis is less robust. Men are more likely to endorse the item than women matched on AS, though overall endorsement is low. The result may stem from stronger affiliation with work in males (Bleustein, 2001; Bleustein & Noumair, 1996). Males may view the reference to 'task' in item 2 as an indication of competency-related work activity, making them unlikely to endorse the item unless they are particularly sensitive to anxiety.

Limitations of this study do deserve comment. Although the ethnic and racial distribution of the sample was comparable to that of the region, the sample was primarily Caucasian, which limits generalizability of the findings to other ethnic groups. The approach of collapsing individual minority groups into one sample is also a limitation. We are well aware of the heterogeneity of the individuals in this group and apologize for taking an approach that overlooks obvious differences. We did this for reasons of statistical power. The sample was primarily composed of college students with a mean age of approximately 19 years. These findings may not generalize to the population as a whole. Future studies need to investigate ethnic and age differences on the AS construct. The low rates of endorsement for item 2 deserve future investigation as well. In a sample where the rate of endorsement was higher, perhaps gender bias would be more evident and would make a difference in total scale scores.

Gender bias has important implications for past and future work with the ASI and with anxiety in general. The impact of gender bias on previously reported findings with the ASI may need to be re-examined. In the future, researchers may want to determine whether conclusions are altered by leaving biased items out of the ASI total score. More investigation of the ASI both with and without the items in question is warranted. In this case, the statistical difference in the ASI across gender was due to measurement bias rather than a true effect. It is extremely important, given that females are twice as likely as males to receive a diagnosis of an anxiety disorder (American Psychiatric Association, 2000), to examine whether differences could be the result of measurement bias instead of actual phenomena.

Acknowledgements

Our hearty thanks to Elana Gordis and the members of the Anxiety Disorders Research Program. A grant awarded to the third author from the National Institute of Mental Health (MH6010701) helped support this work.

References

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders* (4th ed.). Text Revision. Washington, DC: American Psychiatric Association.

Bleustein, D. L. (2001). Work, masculinity, and social class: learning to walk like a man. *Society for the Psychological Study of Men and Masculinity Bulletin, 6*(1), 9–10.

Bleustein, D. L., & Noumair, D. A. (1996). Self and identity in career development: implications for theory and practice. *Journal of Counseling & Development, 74*, 433–441.

Brown, T. A., DiNardo, P. A., & Barlow, D. H. (1994). *Anxiety disorders interview schedule for DSM-IV*. New York: Oxford University Press.

Camilli, G., & Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational and Behavioral Statistics, 24*(4), 323–341.

Deacon, B. J., & Abramowitz, J. S. (2006). Anxiety sensitivity and its dimensions across the anxiety disorders. *Journal of Anxiety Disorders, 20*, 837–857.

Earleywine, M. (2006). Schizotypy, marijuana, and differential item functioning. *Human Psychopharmacology Clinical and Experimental, 21*, 455–461.

Gelin, M. N., & Zumbo, B. D. (2003). Differential item functioning results may change depending on how an item is scored: an illustration with the center for epidemiologic studies depression scale. *Educational and Psychological Measurement, 63*, 65–74.

Jang, K. L., Stein, M. B., Taylor, S., & Livesley, W. J. (1999). Gender differences in the etiology of anxiety sensitivity: a twin study. *Journal of Gender-Specific Medicine, 2*(2), 39–44.

Liu, I., & Agresti, A. (1996). Mantel–Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics, 52*(4), 1223–1234.

Fidalgo, A. M., Ferreres, D., & Muñiz, J. (2004). Utility of the Mantel–Haenszel procedure for detecting differential item functioning in small samples. *Educational and Psychological Measurement, 64*(6), 925–936.

Maller, R. G., & Reiss, S. (1992). Anxiety sensitivity in 1984 and panic attacks in 1987. *Journal of Anxiety Disorders, 6*, 241–247.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel–Haenszel statistic. *Educational and Psychological Measurement, 52*, 443–451.

McNally, R. J. (2002). Anxiety sensitivity and panic disorder. *Biological Psychiatry, 52*, 938–946.

O'Neil, J. M., Helms, B., Gable, R., David, L., & Wrightsman, L. (1986). Gender role conflict scale: college men's fear of femininity. *Sex Roles, 14*, 335–350.

Penfield, R. D. (2007a). An approach for categorizing DIF in polytomous items. *Applied Measurement in Education, 20*(3), 335–355.

Penfield, R. D. (2007b). *DIFAS 4.0 Differential item functioning analysis system: user's manual*. unpublished manuscript.

Peterson, R. A., & Reiss, S. (1993). *Anxiety Sensitivity Index revised test manual*. Worthington, OH: IDS Publishing Corporation.

Radloff, L. S. (1977). The CES-D scale: a self-report depression scale for research in the general population. *Applied Psychological Methods, 3*, 385–401.

Reiss, S., Peterson, R. A., Gursky, D. M., & McNally, R. J. (1986). Anxiety sensitivity, anxiety frequency, and the prediction of fearfulness. *Behaviour Research and Therapy, 20*, 135–152.

Savage, D. D., Corwin, L., McGee, D. L., Kannel, W. B., & Wolf, P. A. (1985). Epidemiologic features of isolated syncope: the Framingham study. *Stroke, 16*(4), 626–629.

Schaub, M., & Williams, C. (2007). Examining the relations between masculine gender role conflict and men's expectations about counseling. *Psychology of Men & Masculinity, 8*(1), 40–52.

Stewart, S. H. & Baker, J. M. (1999). Gender differences in anxiety sensitivity. *Anxiety Disorders Association of America Reporter, 10*(3), 1, 17–20.

Stewart, S. H., Taylor, S., & Baker, J. M. (1997). Gender differences in dimensions of anxiety sensitivity. *Journal of Anxiety Disorders, 11*(2), 179–200.

Vujanovic, A. A., Arrindell, W. A., Bernstein, W. A., Norton, P. J., & Zvolensky, M. J. (2007). Sixteen-item Anxiety Sensitivity Index: confirmatory factor analytic evidence, internal consistency, and construct validity in a young adult sample from the Netherlands. *Assessment, 14*(2), 129–143.

Zinbarg, R. E., Mohlman, J., & Hong, N. N. (1999). *Dimensions of anxiety sensitivity. Anxiety sensitivity: theory, research and treatment of the fear of anxiety*. Mahwah, NJ: Lawrence Erlbaum. pp. 83–114.

Please cite this article in press as: Van Dam, N.T, et al., Gender bias in the sixteen-item Anxiety Sensitivity Index: An application of polytomous differential item functioning. J Anxiety Disord (2008), doi:10.1016/j.janxdis.2008.07.008