Published in final edited form as:

J Psychopathol Behav Assess. 2015 June ; 37(2): 306–317. doi:10.1007/s10862-014-9455-9.

Assessing the Straightforwardly-Worded Brief Fear of Negative Evaluation Scale for Differential Item Functioning Across Gender and Ethnicity

Jared K. Harpole,
Department of Psychology, University of Kansas, Lawrence, KS, USA

Cheri A. Levinson,
Department of Psychology, Washington University in St. Louis, St. Louis, MO, USA

Carol M. Woods,
Department of Psychology, University of Kansas, Lawrence, KS, USA

Thomas L. Rodebaugh,
Department of Psychology, Washington University in St. Louis, St. Louis, MO, USA

Justin W. Weeks,
Department of Psychology, Ohio University, Athens, OH, USA

Patrick J. Brown,
Department of Psychology, Washington University in St. Louis, St. Louis, MO, USA

Department of Psychiatry, Columbia University College of Physicians and Surgeons, New York, NY, USA

Richard G. Heimberg,
Department of Psychology, Temple University, Philadelphia, PA, USA

Andrew R. Menatti,
Department of Psychology, Ohio University, Athens, OH, USA

Carlos Blanco,
Department of Psychiatry, Columbia University College of Physicians and Surgeons, New York, NY, USA

Anxiety Disorders Clinic, New York State Psychiatric Institute, New York, NY, USA

Franklin Schneier, and
Department of Psychiatry, Columbia University College of Physicians and Surgeons, New York, NY, USA

Anxiety Disorders Clinic, New York State Psychiatric Institute, New York, NY, USA

Correspondence to: Jared K. Harpole, jared.harpole@gmail.com.

Conflict of Interest Jared K. Harpole, Cheri A. Levinson, Carol M. Woods, Thomas L. Rodebaugh, Justin W. Weeks, Patrick J. Brown, Richard G. Heimberg, Andrew R. Menatti, Carlos Blanco, Franklin Schneier, and Michael Liebowitz all declare that there were no conflicts of interest.

Experiment Participants All data sets that were collected in this study had Institutional Review Board approval and all participants gave informed consent before entering the studies.

Michael Liebowitz
Department of Psychiatry, Columbia University College of Physicians and Surgeons, New York, NY, USA

Jared K. Harpole: jared.harpole@gmail.com

Abstract

The Brief Fear of Negative Evaluation Scale (BFNE; Leary *Personality and Social Psychology Bulletin*, 9, 371–375, 1983) assesses fear and worry about receiving negative evaluation from others. Rodebaugh et al. *Psychological Assessment, 16*, 169–181, (2004) found that the BFNE is composed of a reverse-worded factor (BFNE-R) and straightforwardly-worded factor (BFNE-S). Further, they found the BFNE-S to have better psychometric properties and provide more information than the BFNE-R. Currently there is a lack of research regarding the measurement invariance of the BFNE-S across gender and ethnicity with respect to item thresholds. The present study uses item response theory (IRT) to test the BFNE-S for differential item functioning (DIF) related to gender and ethnicity (White, Asian, and Black). Six data sets consisting of clinical, community, and undergraduate participants were utilized (N=2,109). The factor structure of the BFNE-S was confirmed using categorical confirmatory factor analysis, IRT model assumptions were tested, and the BFNE-S was evaluated for DIF. Item nine demonstrated significant non-uniform DIF between White and Black participants. No other items showed significant uniform or non-uniform DIF across gender or ethnicity. Results suggest the BFNE-S can be used reliably with men and women and Asian and White participants. More research is needed to understand the implications of using the BFNE-S with Black participants.

Keywords

Differential item functioning; Measurement invariance; Item response theory; Social anxiety disorder; Fear of negative evaluation

---

The Brief Fear of Negative Evaluation Scale (BFNE; Leary 1983) [a shortened version of the Fear of Negative Evaluation Scale (FNE; Watson and Friend 1969)] assesses fear and worry about receiving negative evaluation from others. Fear of negative evaluation is theorized to be a core feature of social anxiety disorder (Haikal and Hong 2010; Heimberg et al. 2010), and the BFNE is often used in studies of social anxiety disorder and of disorders and problems that may have a social evaluative component. For example, the BFNE has been utilized to measure an aspect of social anxiety in populations with eating disorders (e.g., Gilbert and Meyer 2005; Levinson and Rodebaugh 2012), schizophrenia (Blanchard et al. 1998), problem drinking (Lewis and O'Neill 2000), depression (O'Connor et al. 2002), and body dysmorphic disorder (Zimmerman and Mattia 1998), as well as in patients undergoing bariatric surgery (Adams et al. 2011). Thus, there is much research showing the wide clinical and research utility of the BFNE.

Research on the psychometric properties of the BFNE is promising and suggests that the brief version captures more information than the full version (e.g. FNE) of the scale (Rodebaugh et al. 2004). Confirmatory factor analyses show that the BFNE has a 2-factor solution (a straightforwardly-worded factor and a reverse-worded factor) that exhibits

excellent fit and that the straightforwardly-worded factor is better able to predict social anxiety than the reverse-worded factor in both samples of undergraduate and persons with anxiety disorders (Carleton et al. 2006; Carleton et al. 2011; Rodebaugh et al. 2004; Weeks et al. 2005). Item response theory analyses demonstrate that the BFNE provides information across a wide range of severity levels of the latent construct and that the straightforwardly-worded items are associated with higher discrimination parameters than the reverse-worded items (Rodebaugh et al. 2004). Collins et al. (2005) report that the BFNE distinguishes between patients with social anxiety disorder and panic disorder; in addition, in patients with social anxiety disorder, the straightforwardly-worded items predict unique variance in social anxiety (whereas the reverse-worded items do not) (Weeks et al. 2005). Overall, this research provides support for use of the straightforwardly-worded items of the BFNE.

Fear of negative evaluation and constructs related to social anxiety have been studied in both genders and ethnic groups. Okazaki (1997) found that Asian-Americans scored higher on the FNE scale than did Caucasian Americans. Relatedly, Hambrick et al. (2010) found that African-American and Asian-American undergraduates responded differently than did Caucasian undergraduates on two commonly used measures of social anxiety and worry (Social Interaction Anxiety Scale, Mattick and Clarke 1998; Penn State Worry Questionnaire, Meyer et al. 1990). The BFNE has also been utilized in both genders and many ethnic groups (see Carleton et al. 2011; Norton and Weeks 2009). However, research on the BFNE related to threshold invariance across ethnic and gender groups is lacking.

Norton and Weeks (2009) tested the measurement invariance of the BFNE (using only the straightforwardly-worded items; BFNE-S) and the Fear of Positive Evaluation Scale using both classical factor analysis and categorical confirmatory factor analysis in African American, Asian, Caucasian, and Hispanic undergraduate samples. They found no difference in factor loadings or latent variances and covariance between samples. However, they did not test for invariance of the threshold parameters across groups. Further, Carleton et al. (2011) tested the BFNE for metric invariance related to gender and found that responses were similar across gender (with the straightforwardly-worded items exhibiting fewer differences when compared to the reverse-worded items). Like Norton and Weeks (2009) Carleton et al. (2011) did not test the BFNE for threshold invariance. When measurement non-invariance is present in the threshold parameters, item responses will be universally biased either higher or lower across the range of the latent variable based on group membership. This leads to certain groups to obtaining higher or lower scores on a measure based on group membership and not on the latent construct.

Of the aforementioned studies evaluating the BFNE, only Norton and Weeks (2009) used appropriate methods for addressing the ordinal nature of the data. When classical factor analysis is applied to ordinal outcomes such as those on the BFNE, this can lead to biased parameter estimates and potentially alter the factor structure across groups (Lubke and Muthén 2004; Wirth and Edwards 2007). A modern framework to test for measurement invariance across groups with categorical outcomes is item response theory (IRT). IRT is an analytical technique using latent variable models for analyzing categorical data (see e.g., De Ayala 2009). An example of a two-parameter logistic (2PL) IRT model for a dichotomous item is given by

$$P(y=1|\theta)=\frac{1}{1+\exp[-a(\theta-b)]} \quad (1)$$

where *y* is the item response, *a* is the discrimination parameter, θ is the latent construct, and *b* is the threshold or item difficulty parameter. The *a* parameter in Eq. 1 indicates how well the item discriminates between individuals with low and high levels of the latent construct. Theta indicates the level of the latent construct for a given individual and *b* indicates the level of the latent construct an individual must have before the probability of responding in category zero versus one is 50 %.

The practice of testing for measurement invariance within IRT involves testing for *differential item functioning* (DIF). DIF occurs when groups are matched on the latent variable and the probability of responding in a specific category is different for group A versus group B on a given item (Thissen et al. 1986; Thissen et al. 1993). There are two types of DIF, uniform and non-uniform. Uniform DIF affects the threshold parameter and indicates that the probability of responding in a given category is different for group A versus B across the range of the latent variable. Non-uniform DIF influences both the threshold and discrimination parameters and indicates that the probability of responding in a given category is different in group A versus B but changes across the range of theta. To test for DIF, groups must be linked on the latent variable to establish a common scale. In the present study, anchor items were empirically selected and used to link the metric across groups for DIF testing. Anchor items are those deemed least likely to exhibit DIF and therefore presumed DIF-free in the analysis.

It is important to assess both uniform and non-uniform DIF of an assessment to ensure that the scale is a valid measure of the construct (Millsap 2011). The purpose of the present research was to analyze the BFNE scale for both uniform and non-uniform DIF related to gender and ethnicity using an IRT framework to take into account the ordinal nature of the items. Our study extends previous research in several ways. First, we tested both uniform and non-uniform DIF whereas both Carleton et al. (2011) and Norton and Weeks (2009) tested only non-uniform DIF. To the best of our knowledge the present study is the first to examine both non-uniform and uniform DIF on the BFNE-S. Second we used appropriate methods for the scale of measurement (ordinal versus continuous) which extends the work by Carleton et al. (2011). Third, our study consisted of a more heterogeneous sample compared with Norton and Weeks (2009). In Norton and Weeks (2009) an undergraduate sample was used to conduct all analyses. Although this sample did include 4 different ethnic groups, it was still limited to undergraduate students at the University of Houston. Our study included undergraduate, community, and clinical participants from multiple US sites. Fourth, we opted to test the BFNE-S as a stand-alone scale so as to prevent potential influence from additional latent variables. Norton and Weeks (2009) utilized a different model to test for measurement invariance of the BFNE-S versus the one considered here. Taken in combination these four points extend work by Carleton et al. (2011) and Norton and Weeks (2009).

These analyses were conducted using a combination of six data sets with a total of 2,109 participants. First, the factor structure of the BFNE was confirmed and the assumptions of the IRT model were tested. Next, anchor items were selected for the DIF analysis using an iterative purification method. Then these anchor items were used to test for DIF across gender and ethnicity. The results of these analyses should provide a version of the BFNE-S that can be used with diverse clinical populations.

Methods

Participants

A combination of six datasets that included the BFNE consisting of clinical, community, and undergraduate participants (N=2,253) was utilized. Of the 2,253 participants, 25 (1.11 %) were deleted because they were missing values on gender, ethnicity, or both. Further, there were five ethnicity categories with very few participants (American Indian, n=11, Hispanic, n=27, Multi-ethnic, n=32, Not Listed, n=48, and Caribbean, n=1) that could not be used in the analysis (119, 5.28 %). For the remaining 2,109 participants, the average age was 30.56 (SD=20.18); 1,395 were women (66.15 %) and 714 men (33.85 %); 166 (7.87 %) were Black, 173 (8.02 %) Asian, and 1,770 (83.92 %) White. The demographic characteristics of the six individual data sets are described in detail below.

Rodebaugh et al. (2011) reported on a significant subset of these participants, and additional studies, noted below, employed some measures from some of these datasets; however, no previous study examined these data in regard to DIF of the BFNE. A description of each dataset follows. Participants from Dataset 1 included 61 individuals with generalized social anxiety disorder (GSAD) as determined by two structured interviews (n=27) and participants who displayed no evidence of GSAD on the same interviews (NOSAD, n=24). There was also a subset of participants who did not have GSAD, but had social anxiety that was not low enough to be considered NOSAD (e.g., evidence of specific social anxiety disorder) (n=10). Participants were 46 (75.40 %) White, 14 (23.00 %) Black and one (1.60 %) Asian; 35 (57 %) were females, and the median age of participants was 34.98 years (SD=11.86). Participants with GSAD were recruited through advertisement of the study online and via flyers posted in public and at clinics in a Midwest metropolitan area. Participants with NOSAD were selectively recruited from a volunteer registry to be demographically equivalent to the GSAD group. Participants were excluded if they were currently psychotic, manic, or acutely suicidal as assessed by structured clinical interview, or displayed any other psychological problem in need of immediate treatment. The majority of participants in Dataset 1 completed a prisoner's dilemma task described in Rodebaugh et al. (2013).

Participants in Dataset 2 (n=45) were recruited by the same laboratory as Dataset 1 for a study of relationships in individuals with GSAD and included individuals diagnosed with GSAD (n=26) via the Structured Clinical Interview for DSM IV Axis I Disorders (SCID; First et al. 1996) in conjunction with the clinician-administered Liebowitz Social Anxiety Scale (LSAS; Liebowitz 1987) and participants who displayed no evidence of social anxiety disorder (NOSAD) (n=19) based on the same interview. Participants were 30 (88.36 %) White, 13 (7.22 %) Black, and 2 (4.44 %) Asian; 35 (78.00 %) were female, and the mean age was 36.51 (SD= 13.94). Recruitment procedures and inclusion/exclusion criteria were

similar to those for Dataset 1. Data collection for this project was in progress when this study was conducted; papers regarding these participants will be forthcoming but will not focus on DIF in the BFNE.

Participants in Dataset 3 consisted of 180 adult patients who were recruited for participation in a treatment study in one of two Northeastern cities from 2005 to 2007. Of these participants, 172 were diagnosed with GSAD using either the SCID or the Anxiety Disorder Interview Schedule for DSM-IV, Lifetime version (ADIS-IV-L; DiNardo et al. 1994) and the remainder (n=8) had a current diagnosis of non-generalized social anxiety disorder. Participants were 114 (63.34 %) White, 44 (24.44 %) Black, and 22 (12.22 %) Asian; 73 (41.00 %) were female; the mean age was 32.35 (SD=11.86). Most participants were recruited from primary care offices, mental health practices, or were self-referred from advertisements. Most participants took part in a study concerning augmentation of medication treatment with cognitive behavioral therapy; findings from this study will be forthcoming. This sample also overlaps partially, but not fully, with that of Weeks et al. (2012). That study provided some information about the BFNE factor structure, but focused on a separate measure and did not examine DIF.

Participants from Dataset 4 included 472 adults from a Midwestern metropolitan community who were recruited through community volunteer registries. Participants were 438 (92.80 %) White, 27 (5.70 %) Black, and 7 (1.50 %) Asian; 333 (71.00 %) were female; the mean age was 61.43 (SD=19.49). These data were collected between 2007 and 2008. This sample has also been reported on by Rodebaugh et al. (2011) and Brown and Roose (2011), but none of the results overlap with those presented here.

Participants from Dataset 5 were 463 undergraduates who completed a questionnaire packet to receive credit as part of their coursework at a private Midwestern metropolitan university in the same community as Dataset 4. Participants were 318 (68.69 %) White, 33 (7.12 %) Black, and 112 (24.19 %) Asian; 318 (69.00 %) were female; the mean age was 19.04 (SD=1.05). Parts of these data have been reported in several studies, but none have focused on the item properties of the BFNE (e.g., Levinson and Rodebaugh 2011). These data were collected in 2007 and 2008.

Participants from Dataset 6 were 888 undergraduates from a public Midwestern university (not the same as Dataset 5). Participants were 824 (92.79 %) White, 35 (3.94 %) Black, and 29 (3.27 %) Asian; 601 (68.00 %) were female; the mean age was 19.08 (SD=1.57). Participants were recruited from an introductory psychology class and completed all measures online. Parts of these data have been reported in Levinson et al. (2013).

Measures

BFNE—The BFNE is a self-report questionnaire developed to assess participants' fear of negative evaluation (Leary 1983). The BFNE was based on the 30-item Fear of Negative Evaluation Scale (FNE; Watson and Friend 1969). Participants are asked to indicate how characteristic each of the 12 statements is of them on a 1–5 Likert-type scale. Items two, four, seven, and 10 are reverse-worded items. In an undergraduate sample, coefficient alpha

and 4-week test-retest reliability have been reported to be 0.90 and 0.75, respectively, for the total scale (Leary 1983).

BFNE Straightforwardly-worded Items—Rodebaugh et al. (2004) and Weeks et al. (2005) showed that the straightforwardly-worded items (BFNE-S) and reverse-worded items (BFNE-R) comprised two separate factors. Furthermore, they recommended that the BFNE-S be used instead of the BFNE-R due to its superior psychometric properties. The BFNE-S is composed of items one, three, five, six, eight, nine, 11, and 12 from the original BFNE. Several studies report the BFNE-S had an α>0.92 in undergraduate (Rodebaugh et al. 2004) and clinical samples (Carleton et al. 2011; Weeks et al. 2005). Carleton et al. (2011) conducted a review of three different ways to deal with the inadequacy of the BFNE-R, and their findings indicated that the original eight-item BFNE-S (omitting the BFNE-R items) performed best. For these reasons, DIF analyses focus only on the BFNE-S in the present study.

Procedure

Data Analysis—First, a categorical confirmatory factor analysis (CCFA) was performed on the BFNE using the factor structure suggested by Rodebaugh et al. (2004). Second, overall IRT model assumptions were assessed and anchor items were empirically selected for the BFNE-S. Third, the BFNE-S was tested for DIF using a version of Lord's (1980) χ^2 test implemented in flexMIRT™software (version 1.88; Cai 2012) that was recently improved because it uses concurrent linking and more accurately estimated standard errors (Cai 2008; Cai et al. 2013; Langer 2008; Woods et al. 2013).

Categorical Confirmatory Factor Analyses—The 2,109 participants described above were used in the categorical confirmatory factor analysis (CCFA) and multiple group CCFA (MGCCFA). The average age was 30.56 (*SD*=20.18); 1,395 were women (66.15 %) and 714 men (33.85 %). The MGCCFA was used to assess invariance across sites prior to pooling the samples, and the CCFA was used to confirm the factor structure of the BFNE once site invariance was established. There were 33 cases (1.47 %) of the 2,109 with at least one missing value. To avoid deleting these 33 cases in the CCFA and MGCCFA, multiple imputation (MI; Rubin 1987) was used; 20 imputed data sets were created using the Amelia II package in R (version 3.02) (Honaker et al. 2011; R Core Team 2013). These imputed data sets were then used in Mplus (version 7.0; Muthén and Muthén 2012) to conduct the CCFA by combining the imputed data sets (Rubin 1987; Schaefer and Olsen 1998).

Because six data sets from different sites were being pooled, measurement invariance across sites was tested to ensure that the BFNE was measuring the same construct across sites. For purposes of the invariance testing, Dataset 1 (n=68) and Dataset 2 (n=48) were combined to obtain a larger sample to improve parameter estimation. Both Datasets 1 and 2 were collected from the same lab and both consisted of participants with and without GSAD. No other data sets were combined; thus invariance testing was carried out with five sites.

A two-factor model was fitted in each site with eight items (items 1, 3, 5, 6, 8, 9, 11, and 12) loading on a straightforwardly-worded factor and four items loading on a reverse-worded factor (items 2, 4, 7, and 10) (Rodebaugh et al. 2004). We used the weighted least squares

with mean and variance adjustments (WLSMV) estimator with polychoric correlations in Mplus (see Muthén and Muthén 2012, for details). Configural, weak, and strong invariance were tested across sites using a MGCCFA model. Three fit indices were used to assess global fit: (1) the Tucker and Lewis Index (TLI; Tucker and Lewis 1973), (2) Bentler's (1990) Comparative Fit Index (CFI), and (3) the Root Mean Square Error of Approximation (RMSEA; Steiger and Lind 1980). Cut-off criteria for these measures were obtained from the recommendations of Hu and Bentler (1998; 1999) and MacCallum et al. (1996) ($RMSEA<0.08$, CFI and $TLI>0.95$).

Given the problematic behavior of the χ^2 deviance test in large samples, the χ^2 deviance test was not used for invariance testing (Cheung and Rensvold 2002; Elosua 2011; Muthén and Muthén 2012). The change in Bentler's (1990) Comparative Fit Index (CFI) was used instead. A change in CFI of 0.01 or less is indicative of measurement invariance when comparing two nested models. The change in CFI has been shown to perform well in CCFA (see Elosua 2011) and may overcome the problematic behavior of the χ^2 deviance test in large samples (Cheung and Rensvold 2002). After site invariance was established, a CCFA was fitted using the same two-factor model described above to confirm the factor structure. The factor structure of the CCFA was evaluated using the same global fit indices and cut-off criteria mentioned previously.

Item Response Theory Analysis—IRT was carried out using flexMIRT™ (version 1.88; Cai 2012) using data from the 2,109 participants described previously. Because BFNE-S response options consist of five ordered categories, the graded response model (GRM; Samejima 1969) was used. The GRM is given by

$$P(x_{ijk} \geq k|\theta) = \frac{1}{1+\exp[-a_j(\theta - b_{ijk})]}, \quad (2)$$

for k=0, 1, 2, …., K_j; j=0, 1, 2, …., J; and i=1, 2, …., I. Note that x is the response to item j, k is the number of response categories with K_j greater than two, J is the number of items, and I is the number of individuals. There are k-1 thresholds in any given item. The GRM models the probability that individual i chose category k or higher for item j.

All IRT models were estimated using marginal maximum likelihood (Bock and Aitkin 1981) and identified by setting the latent mean and variance to zero and one, respectively. Initially a unidimensional IRT model was fitted to the BFNE-S factor to assess the global and item-level fit and to check for local independence. The assumption of local independence in IRT indicates that once a person's level of latent variable is taken into account, all item responses are statistically independent (see De Ayala 2009, p. 20). Incomplete item responses were handled by the estimation methods in flexMIRT™ so the full data set including incomplete responses was used.

Global model fit was assessed using the M_2 statistic and RMSEA. The M_2 statistic uses information from the bivariate sample moments to compute a statistic that is asymptotically Chi-square distributed (Maydeu-Olivares and Joe 2005, 2006). However, the M_2 statistic is

often too powerful when sample sizes are large. Thus, the RMSEA was also computed to assess global model fit to mitigate the effects of sample size on model fit. Goodness of fit cut-offs were evaluated with the same criteria as the MGCCFA and CCFA. Orlando and Thissen's (2000, 2003) S-χ^2 item level fit statistics were calculated to assess item level fit. Nonsignificant S-χ^2 statistics suggest good fit. Additionally item level fit was evaluated graphically using MODFIT 3.0 (Stark 2008). The assumption of local dependence (LD) was evaluated using the Chen and Thissen (1997) LD statistics implemented in flexMIRT™. These statistics use the observed and expected item response frequencies in two by two contingency tables for all item pairs and are approximately chi-square distributed with one degree of freedom. Any LD statistics that are larger than 10 are cause for concern (Cai et al. 2013) and were flagged for potential LD.

Differential Item Functioning—For each data set, anchor items were empirically selected using a procedure proposed by Woods (2009a). Previous research suggests that the anchor set should be approximately 10 % and 20 % of the total number of test items (Wang et al. 2009; Woods 2009a, b); thus, two anchor items (25 % of the test length) were chosen for each analysis. DIF analyses were conducted using the *test candidate items* option in flexMIRT™. In this approach, the latent mean and variance for the reference group are fixed to zero and one, respectively, while the latent mean(s) and variance(s) for the focal group(s) are estimated. The algorithm provides an overall (omnibus) χ^2 test of DIF in any item parameter, as well as conditional χ^2 tests of uniform and non-uniform DIF.

For the ethnic group comparisons, a contrast matrix was used to compare White participants with Asian participants and White participants with Black participants. The contrast matrix was

$$\begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}, \quad (3)$$

with 1 s for the White reference group and −1 s for the focal groups. To help control the Type I error rate for both gender and ethnicity DIF tests, the Benjamini and Hochberg (1995) procedure (BH) was applied to all tests (Thissen et al. 2002). For each set of tests (omnibus, discrimination parameters, and threshold parameters) the alpha level was corrected within that set of tests. The procedure was carried out in R (v. 3.02) using the *p.adjust* function in the stats package to calculate adjusted p-values for the χ^2 tests for DIF (R Core Team 2013).

Results

MGCCFA

The results for the MGCCFA indicated that the BFNE was invariant across sites so the data sets could be pooled. The configural model showed good fit ($\chi^2_{(297)}=1124.23$, $RMSEA$=0.08, CFI=0.984, TLI=0.982) which established a common factor pattern across the five sites. The weak invariance and strong invariance models also showed good fit

($\chi^2_{(337)}$=1098.16, *RMSEA*=0.07, *CFI*=0.985, *TLI*=0.985, Δ *CFI* = −0.001) and ($\chi^2_{(489)}$=1345.53, *RMSEA* = 0.06, *CFI* = 0.983, *TLI* = 0.99, Δ*CFI* = −0.002), respectively.

CCFA

For the BFNE, the model fit indices from the two-dimensional model suggested good fit [*TLI* = 0.99, *CFI* = 0.99, *RMSEA* = 0.07, $\chi^2_{(53)}$=604.03, p<0.001], confirming the factor structure proposed by Rodebaugh et al. (2004). All standardized factor loadings were statistically significant. For the BFNE-S the magnitudes of the standardized loadings were between 0.83 and 0.91 and for the BFNE-R the standardized loadings were between 0.67 and 0.80. The between factor correlation between BFNE-S and BFNE-R was statistically significant (r = −0.495, p<0.001). Standardized factor loadings are reported in Table 1. As planned, the BFNE-S was the subject of further tests described below.

Global IRT

Results for the IRT model on the BFNE-S indicated adequate fit (M_2 = 1,745.95, df = 440, p<0.001, and *RMSEA* = 0.04). The RMSEA value of 0.04 indicates that the model has good global fit according to the criteria given by Hu and Bentler (1998; 1999) and MacCallum et al. (1996). A Bonferroni correction was used to assess the significance of the S-χ^2 by using 0.05/8 = 0.006 as the critical value. Item fit was adequate for items one (S-X^2=105.60, df=84, p>0.006), three (S-χ^2 =113.60, df=80, p>0.006), five (S-χ^2 =92.50, df=70, p>0.006), eight (S-χ^2 =91.70, df=76, p>0.05), nine (S-χ^2 = 106.00, df=75, p>0.006), and 12 (S-χ^2 =101.10, df=75, p>0.006). Items six (S-χ^2 =119.20, df=69, p<0.006) and 11 (S-χ^2 =134.70, df=83, p<0.006) fit significantly differently than predicted by the model. Examination of the item fit plots using MODFIT (v. 3.0) indicated that both items six and 11 may have some item level misfit.

The LD tests indicated that item pairs five and six (*LD*= 36.30), 11 and one (*LD*=10.20), 12 and five (*LD*=12.20), 12 and six (*LD*=12.20), 11 and 12 (*LD*=10.00), nine and six (*LD*=15.40), and eight and nine (*LD*=17.20) had χ^2 greater than or equal to 10 indicating that these items may exhibit local dependence. Work by Harpole and Woods (2013; 2014) showed that violations of LD caused by similarly worded items did not adversely affect Type I errors or power when testing for DIF unless all items in the anchor set were contaminated and the degree of LD was large. To ensure control over Type I errors and power in the presence of LD anchors were selected based on Woods (2009a), while ensuring that at least one of the two anchors was not flagged for LD.

DIF Analysis of BFNE-S for Gender and Ethnicity

Using the anchor selection method from Woods (2009a) and the constraint of at least one LD free anchor, two anchors were selected for both gender (items three and 11) and ethnicity (items one and three). The anchor items were different for ethnicity and gender to ensure that at least one of the two anchors was not part of an LD pair.[1] This avoids the Type I error inflation and power reduction problems demonstrated in Harpole and Woods (2013; 2014) when testing for DIF. IRT parameter estimates by gender (men and women) are presented in Table 2. The latent mean and variance for women were fixed to zero and one,

respectively; the latent mean and variance for men were estimated [0.10 (*SE*=0.08) and 1.06 (*SE*=0.14), respectively]. Results of DIF testing for gender are presented in Table 3. The results indicated that none of the items functioned differentially for men versus women after controlling for true mean differences on fear of negative evaluation.

Item parameter estimates by ethnicity (White, Asian, and Black) are presented in Table 4. The latent mean and variance for White participants were fixed to zero and one, respectively. The latent mean and variance for the Asian participants relative to White participants were 0.34 (*SE*=0.14) and 0.57 (*SE*= 0.09) respectively; the latent mean and variance for Black participants relative to White participants were 0.16 (*SE*= 0.08) and 1.58 (*SE*=0.27) respectively. Results of DIF testing for ethnicity are presented in Table 5. Item nine showed significant ($\chi^2_{a(1)}=19.50$, *BH corrected p*<0.01) non-uniform DIF indicating that item nine was more discriminating for White participants (*a* = 3.28) than Black participants (*a* = 2.04). All other items did not show significant uniform or non-uniform DIF.

Figure 1 provides a graphical illustration of the effect size of the non-uniform DIF between White and Black participants (Steinberg and Thissen 2006). Three plots are included: a test characteristic curve (TCC) (i.e. the relation between the BFNE-S in the IRT metric and raw metric) for all eight BFNE-S items, a TCC for all BFNE-S items with item nine (non-uniform DIF item) removed, and the item characteristic curve (ICC) (i.e. relation between an item in the IRT metric and raw metric) for item nine. Each figure plots the expected score function against the level of fear of negative evaluation for a set of items (TCC) or a single item (ICC). The differences between Black and White participants were small for the full scale, with a slight discernible difference between the plots with and without item nine. Nevertheless, when item nine is isolated, it is clear that the expected scores for this item are not equivalent for the two groups. According to the ICC in Fig. 1, below the approximate mean true latent variable score (0.50), the summed score for a Black versus White participant on item nine is expected to be higher even when the individuals are matched on true scores. Above the approximate mean true score (0.50), the bias is in the opposite direction.

Discussion

The purpose of the study was to assess the BFNE-S for uniform and non-uniform DIF in an IRT framework across gender and ethnicity. Results of the DIF analysis indicated that the BFNE-S does not function differentially across gender. This finding extends the work of Norton and Weeks (2009) by showing that the BFNE-S is invariant across gender for both discrimination and threshold parameters. However, the results indicated that item nine (“I am usually worried about what kind of impression I make.”) demonstrated significant non-uniform DIF for White versus Black participants. Item nine was more discriminating for White compared with Black participants.

[1]To test whether the results of the DIF analyses were dependent on anchor items were-ran the analyses for both gender and ethnicity using items three and eight which did not exhibit LD. The results from the DIF analyses for using items three and eight versus the items described above were identical.

Although the non-uniform DIF on item nine was significant and the DIF effect was nontrivial, the TCCs in Fig. 1 show that the impact appears small on the BFNE-S scale as a whole. Thus, the importance of the DIF in item nine depends how the scale is used. If summed scores for the entire scale are used, the BFNE-S is relatively invariant against threats of gender and ethnic groups considered here. However, if item nine is used solely or if a smaller subset of items on the BFNE-S including item nine are used for scoring this may influence the results. If practitioners are attempting to assess actual mean group differences across Black versus White participants in the BFNE-S summed scores, removing item nine it is advisable as small effects could have an impact on the outcome in some circumstances. Further, the finding of item nine exhibiting non-uniform DIF between White and Black participants suggests there could be other items that exhibit DIF among other ethnic groups not tested in the current study.

The present observation of DIF in item nine deviates from Norton and Weeks (2009) who concluded that the factor loadings were invariant. Two possible reasons for this are sample heterogeneity and the model tested. First, the sample from Norton and Weeks (2009) consisted of undergraduates from the University of Houston, whereas our sample consisted of undergraduate, community, and clinical participants from several different regions of the United States. These differences in sample characteristics may have played a role in this finding. Second, Norton and Weeks (2009) used a two factor model with the BFNE-S and the Fear of Positive Evaluation Scale (FPES; Weeks et al. 2008) and the present study only considered the BFNE-S. The influence of the two factor model reported in Norton and Weeks (2009) may have had an influence on failing to find non-invariance in item nine between Black and White participants. These findings provide evidence that the BFNE-S can be used reliably for men and women and for Asian and White participants. However, more research is needed to understand the implications of using the BFNE-S with White and Black participants.

The sample sizes in this study were very unbalanced when assessing DIF for ethnicity (Whites = 1,770, Black = 166, Asian = 173) and power was a concern. Although small focal groups are sometimes combined to increase power for DIF analyses, differences in the parameter estimates were nontrivial for the Asian and Black groups when they were estimated separately (see Table 4) so combining these groups would not have made sense in this case. The sample sizes for the focal groups used in this study are not too small, but a replication of the present study with larger samples of ethnic minority focal groups is indeed warranted.

The findings of this study should be viewed in the context of several limitations. The current sample is large and heterogeneous, yet only a small number of ethnic groups could be evaluated. Further, the sample sizes of the ethnic groups were small and power could have be an issue in detection of other DIF items (e.g. other than item nine for Black and White participants) among Asian, and Black participants compared with White participants. Further research with larger ethnic samples would address this issue. This study pooled six data sets from various sites that consisted of undergraduate, clinical, and community samples. Although measurement invariance of the BFNE-S was demonstrated across sites, there could be regional and demographic differences that we failed to capture in our

analyses. More research is needed to understand the implications of item nine functioning differently between White and Black participants. Further, additional research could assess if the LD found in preliminary IRT analyses has practical implications. Our results should be considered in the context of ethnic and cultural differences that have been shown in the anxiety disorder literature (e.g., Breslau et al. 2006; Smith et al. 2006). For example, some research has suggested that some ethnic groups (African-American) have lower risk of internalizing disorders in general (Smith et al. 2006), whereas other ethnic groups (Asian-American) have a higher risk of internalizing disorders (e.g., Okazaki 1997). We hope that future researchers will consider how measurement across ethnic groups could impact these findings.

Acknowledgments

This research was supported in part by the National Institute of Mental Health (NIMH) grant F31-MH096433-01 to Cheri A. Levinson; National Institute of Health (NIH)/NIMH grant MH090308 to Thomas L. Rodebaugh; NIH grant UL1 RR024992 to Washington University St. Louis; National Institute of Aging (NIA) grant T32 AG0030 and NIMH grant T32 MH20004 to Patrick J. Brown; NIMH grant R01 MH064481-01A1 and GlaxoSmithKline Pharmaceuticals grant 101618 to Richard G. Heimberg; NIH grant K02 DA023200 to Carlos Blanco; NIH grant R01 MH064726 to Michael Liebowitz. Carlos Blanco, Franklin Schneier, and Michael Liebowitz were also supported by the New York Psychiatric Institute. The authors would like to thank the Center for Research Methods and Data Analysis at the University of Kansas for assistance in preparing this manuscript.

References

Adams CE, Myers VH, Barbera BL, Brantley PJ. The role of fear of negative evaluation in predicting depression and quality of life 4 years after bariatric surgery in women. Psychology. 2011; 2:150–154.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B. 1995; 57:289–300.

Bentler PM. Comparative fit indexes in structural models. Psychological Bulletin. 1990; 107:238–246. [PubMed: 2320703]

Blanchard JJ, Mueser KT, Bellack AS. Anhedonia, positive and negative affect, and social functioning in schizophrenia. Schizophrenia Bulletin. 1998; 24:413–424. http://schizophreniabulletin.oxfordjournals.org/. [PubMed: 9718633]

Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika. 1981; 46:443–459.

Breslau J, Aguilar-Gaxiola S, Kendler KS, Su M, Williams D, Kessler RC. Specifying race-ethnic differences in risk for psychiatric disorder in a USA national sample. Psychological Medicine. 2006; 36:57–68. [PubMed: 16202191]

Brown PJ, Roose SP. Age and anxiety and depressive symptoms: the effect on domains of quality of life. International Journal of Geriatric Psychiatry. 2011; 26:1260–1266. [PubMed: 21351152]

Cai L. SEM of another flavor: Two new applications of the supplemented EM algorithm. British Journal of Mathematical and Statistical Psychology. 2008; 61:309–329. [PubMed: 17971266]

Cai, L. flexMIRT™ version 1.88: A numerical engine for multilevel item factor analysis and test scoring. [Computer Software]. Seattle: Vector Psychometric Group; 2012.

Cai, L.; Thissen, D.; du Toit, SHC. IRTPRO:Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Chicago: Scientific Software International; 2013.

Carleton NR, Collimore KC, McCabe RE, Antony MM. Addressing revisions to the brief fear of negative evaluation scale: measuring fear of negative evaluation across anxiety and mood disorders. Journal of Anxiety Disorders. 2011; 25:822–828. [PubMed: 21565463]

Carleton NR, McCreary DR, Norton PJ, Asmundson GG. Brief fear of negative evaluation scale-revised. Depression and Anxiety. 2006; 23:297–303. [PubMed: 16688736]

Chen W-H, Thissen D. Local dependence indexes for item pairs using item response theory. Journal of Educational and Behavioral Statistics. 1997; 22:265–289.

Cheung GW, Rensvold RB. Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling. 2002; 9:233–255.

Collins KA, Westra HA, Dozois DA, Stewart SH. The validity of the brief version of the fear of negative evaluation scale. Journal of Anxiety Disorders. 2005; 19:345–359. [PubMed: 15686861]

De Ayala, RJ. The theory and practice of item response theory. New York, NY: The Guilford Press; 2009. http://www.guilford.com/

DiNardo, PA.; Brown, TA.; Barlow, DH. Anxiety disorders interview schedule for DSM-IV: Lifetime version (ADIS-IV-L). San Antonio: The Psychological Corporation; 1994.

Elosua P. Assessing measurement equivalence in ordered-categorical data. Psicológica. 2011; 32:403–421. http://www.uv.es/revispsi/paraARCHIVES/2011.html.

First, MB.; Spitzer, RL.; Gibbon, M., et al. Structured Clinical Interview for DSM-IV Axis I Disorders (SCID), Clinician Version: Administration Booklet. Washington, DC; American Psychiatric Press; 1996. http://www.appi.org/Home

Gilbert N, Meyer C. Fear of negative evaluation and the development of eating psychopathology: a longitudinal study among nonclinical women. International Journal of Eating Disorders. 2005; 37:307–312. [PubMed: 15856504]

Haikal M, Hong RY. The effects of social evaluation and looming threat on self-attentional biases and social anxiety. Journal of Anxiety Disorders. 2010; 24:345–352. [PubMed: 20176459]

Hambrick JP, Rodebaugh TL, Balsis S, Woods CM, Mendez JL, Heimberg RG. Cross-ethnic measurement equivalence of measures of depression, social anxiety, and worry. Assessment. 2010; 17:155–171. [PubMed: 19915199]

Harpole, JK.; Woods, CM. The impact of local dependence on differential item functioning. Paper presented at the annual meeting of the Psychometric Society; Arnhem, Netherlands. 2013.

Harpole JK, Woods CM. The impact of local dependence on differential item functioning. 2014 Manuscript submitted for publication.

Heimberg, RG.; Brozovich, FA.; Rapee, RM. A cognitive-behavioral model of social anxiety disorder: update and extension. In: Hofmann, SG.; DiBartolo, PM., editors. Social anxiety: Clinical, developmental, and social perspectives. New York: Elsevier; 2010.

Honaker J, King G, Blackwell M. Amelia II: A program for missing data. Journal of Statistical Software. 2011; 45:1–47. http://www.jstatsoft.org.

Hu L, Bentler PM. Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification. Psychological Methods. 1998; 3:424–453.

Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Structural Equation Modeling. 1999; 6:1–55.

Langer, M. Unpublished doctoral dissertation. University of North Carolina at Chapel Hill; 2008. A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation. http://dc.lib.unc.edu/cdm/singleitem/collection/etd/id/2084

Leary MR. A brief version of the fear of negative evaluation scale. Personality and Social Psychology Bulletin. 1983; 9:371–375.

Levinson CA, Rodebaugh TL. Social anxiety and eating disorders: the role of negative social evaluation fears. Eating Behaviors. 2012; 13:27–35. [PubMed: 22177392]

Levinson CA, Rodebaugh TL. Validation of the social appearance anxiety scale: factor, convergent, and divergent validity. Assessment. 2011; 18:350–356. [PubMed: 21467096]

Levinson CA, Rodebaugh TL, Menatti A, Weeks JW. Development and validation of the Social Exercise and Anxiety Measure (SEAM): assessing fears, avoidance, and importance of social exercise. Journal of Psychopathology and Behavioral Assessment. 2013; 35:244–253.

Lewis BA, O'Neill H. Alcohol expectancies and social deficits relating to problem drinking among college students. Addictive Behaviors. 2000; 25:295–299. [PubMed: 10795955]

Liebowitz MR. Social phobia. Modern Problems of Pharmacopsychiatry. 1987; 22:141–173. http://www.karger.com/BookSeries/Home/223929. [PubMed: 2885745]

Lord, FM. Applications of item response theory to practical testing problems. Hillsdale: Lawrence Erlbaum; 1980.

Lubke GH, Muthén BO. Applying multigroup confirmatory factor models for continuous outcomes to likert scale data complicates meaningful group comparisons. Structural Equation Modeling. 2004; 11:514–534.

MacCallum RC, Browne MW, Sugawara HM. Power analysis and determination of sample size for covariance structure modeling. Psychological Methods. 1996; 1:130–149.

Mattick RP, Clarke JC. Development and validation of measures of social phobia, scrutiny, fear, and social interaction anxiety. Behaviour Research and Therapy. 1998; 36:455–470. [PubMed: 9670605]

Maydeu-Olivares A, Joe H. Limited-and full-information estimation and goodness-of-fit testing in 2^n contingency tables. Journal of the American Statistical Association. 2005; 100:1009–1020.

Maydeu-Olivares A, Joe H. Limited information goodness-of-fit testing in multidimensional contingency tables. Psychometrika. 2006; 71:713–732.

Meyer TJ, Miller ML, Metzger RL, Borkovec TD. Development and validation of the Penn State Worry Questionnaire. Behaviour Research and Therapy. 1990; 28:487–495. [PubMed: 2076086]

Millsap, RE. Statistical approaches to measurement invariance. New York: Routledge; 2011. http://www.routledge.com

Muthén, LK.; Muthén, BO. Mplus user's guide. Seventh. Los Angeles, CA: Muthén & Muthén; (1998–2012). http://www.statmodel.com

Norton PJ, Weeks JW. A multi-ethnic examination of socioevaluative fears. Journal of Anxiety Disorders. 2009; 23:904–908. [PubMed: 19560315]

O'Connor LE, Berry JW, Weiss J, Gilbert P. Guilt, fear, submission, and empathy in depression. Journal of Affective Disorders. 2002; 71:19–27. [PubMed: 12167497]

Okazaki S. Sources of ethnic differences between Asian American and White American college students on measures of depression and social anxiety. Journal of Abnormal Psychology. 1997; 106:52–60. [PubMed: 9103717]

Orlando M, Thissen D. Further investigation of the performance of S-χ^2: an item-fit index for use with dichotomous item response theory models. Applied Psychological Measurement. 2003; 27:289–298.

Orlando M, Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models. Applied Psychological Measurement. 2000; 24:50–64.

Core Team R. R: A Language and Environment for Statistical Computing [Computer Software]. Vienna: R Foundation for Statistical Computing; 2013.

Rodebaugh TL, Shumaker EA, Levinson CA, Fernandez KC, Langer JK, Lim MH, Yarkoni T. Interpersonal constraint conferred by generalized social anxiety disorder is evident on a behavioral economics task. Journal of Abnormal Psychology. 2013; 22:39–44. [PubMed: 23231458]

Rodebaugh TL, Woods CM, Thissen DM, Heimberg RG, Chambless DL, Rapee RM. More information from fewer questions: the factor structure and item properties of the original and brief fear of negative evaluation scale. Psychological Assessment. 2004; 16:169–181. [PubMed: 15222813]

Rodebaugh TL, Heimberg RG, Brown PJ, Fernandez KC, Blanco C, Schneier FR, Liebowitz MR. More reasons to be straightforward: Findings and norms for two scales relevant to social anxiety. Journal of Anxiety Disorders. 2011; 25:623–630. [PubMed: 21388781]

Rubin, DB. Multiple imputation for nonresponse in surveys. New York: Wiley; 1987.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Richmond, VA: Psychometric Society; 1969. (Psychometric Monograph No. 17). http://www.psychometrika.org

Schaefer JL, Olsen MK. Multiple imputation for multivariate missing-data problems: a data analysts perspective. Multivariate Behavioral Research. 1998; 33:545–571. [PubMed: 26753828]

Smith SM, Stinson FS, Dawson DA, Goldstein R, Huang B, Grant BF. Race/ethnic differences in the prevalence and co-occurrence of substance use disorders and independent mood and anxiety disorders: results from the national epidemiologic survey on alcohol and related conditions. Psychological Medicine. 2006; 36:987–998. [PubMed: 16650344]

Stark, S. MODFIT: Plot theoretical item response functions and examine the fit of dichotomous and polytomous IRT models to response data [Computer program]. Tampa: Department of Psychology, University of South Florida; 2008.

Steiger, JH.; Lind, J. Paper presented at the Annual Spring Meeting of the Psychometric Society; Iowa City, IA: 1980. Statistically-based tests for the number of common factors. http://www.statpower.net

Steinberg L, Thissen D. Using effect sizes for research reporting: examples using item response theory to analyze differential item functioning. Psychological Methods. 2006; 11:402–415. [PubMed: 17154754]

Thissen D, Steinberg L, Gerrard M. Beyond group-mean differences: the concept of item bias. Psychological Bulletin. 1986; 99:118–128.

Thissen D, Steinberg L, Kuang D. Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. Journal of Educational and Behavioral Statistics. 2002; 27:77–83.

Thissen, D.; Steinberg, L.; Wainer, H. Detection of differential item functioning using the parameters of item response models. In: Holland, PW.; Wainer, H., editors. Differential item functioning. Hillsdale: Lawrence Erlbaum; 1993. p. 67-111.

Tucker LR, Lewis C. A reliability coefficient for maximum likelihood factor analysis. Psychometrika. 1973; 38:1–10.

Wang W-C, Shih C-L, Yang C-C. The MIMIC method with scale purification for detecting differential item functioning. Educational and Psychological Measurement. 2009; 69:713–731.

Watson D, Friend R. Measurement of social-evaluative anxiety. Journal of Consulting and Clinical Psychology. 1969; 33:448–457. [PubMed: 5810590]

Weeks JW, Heimberg RG, Fresco DM, Hart TA, Turk CL, Schneier FR, et al. Empirical validation and psychometric evaluation of the brief fear of negative evaluation scale in patients with social anxiety disorder. Psychological Assessment. 2005; 17:179–190. [PubMed: 16029105]

Weeks JW, Heimberg RG, Rodebaugh TL. The fear of positive evaluation scale: assessing a proposed cognitive component of social anxiety. Journal of Anxiety Disorders. 2008; 22:44–55. [PubMed: 17884328]

Weeks JW, Heimberg RG, Rodebaugh TL, Goldin PR, Gross JJ. Psychometric evaluation of the fear of positive evaluation scale in patients with social anxiety disorder. Psychological Assessment. 2012; 24:301–312. [PubMed: 21966932]

Wirth RJ, Edwards MC. Item factor analysis: current approaches and future directions. Psychological Methods. 2007; 12:58–79. [PubMed: 17402812]

Woods CM. Empirical selection of anchors for tests of differential item functioning. Applied Psychological Measurement. 2009a; 33:42–57.

Woods CM. Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. Multivariate Behavioral Research. 2009b; 44:1–27. [PubMed: 26795105]

Woods CM, Cai L, Wang M. The Langer-improved Wald test for DIF testing with multiple groups: evaluation and comparison to two-group IRT. Educational and Psychological Measurement. 2013; 73:532–547.

Zimmerman M, Mattia JI. Body dysmorphic disorder in psychiatric outpatients: recognition, prevalence, comorbidity, demographic, and clinical correlates. Comprehensive Psychiatry. 1998; 39:265–270. [PubMed: 9777278]

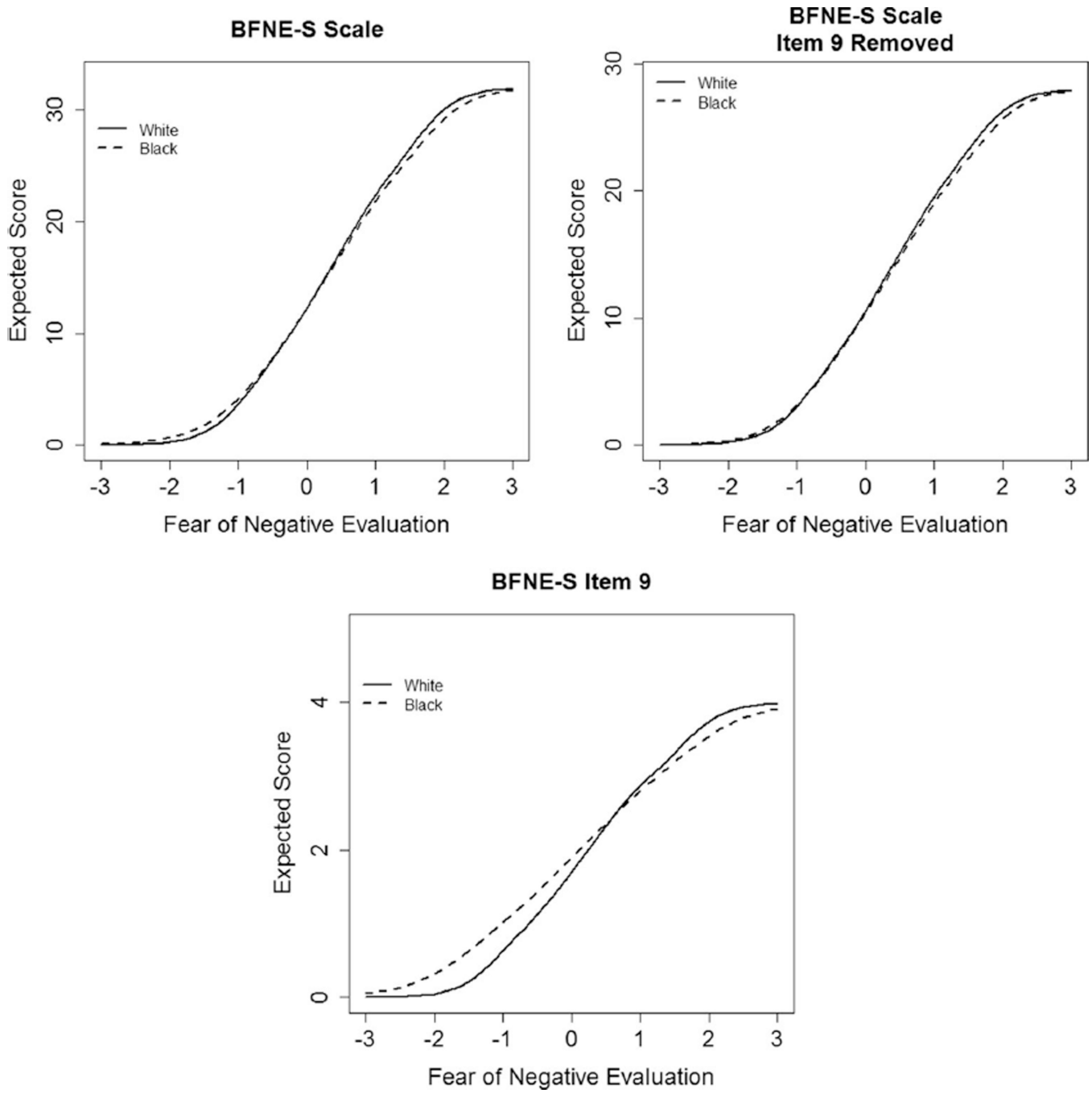

Fig. 1.

Illustration of the Effect Size of Differential Item Functioning with Test and Item Characteristic Curves. Item nine reads “I am usually worried about what kind of impression I make”. This plots the expected score functions against the amount of fear of negative evaluation measured between Black and White participants for the full Brief Fear of Negative Evaluation scale with straightforwardly-worded items (BNFE-S) (*upper left*), the BFNE-S with item nine (non-uniform DIF item) removed (*upper right*), and for item nine

solely (*lower middle*). Expected Score indicates the expected score a participant would obtain according the item response theory model in each group

Table 1

CCFA Parameter Estimates for the BFNE

Item		Standardized Factor Loading	*SE*	Z Statistic	*p*
Straightforwardly-worded factor					
1.	I worry about what people will think of me even when I know it doesn't make any difference.	0.83	0.01	102.43	<0.01
3.	I am frequently afraid of other people noticing my shortcomings.	0.85	0.01	117.01	<0.01
5.	I am afraid that others will not approve of me.	0.90	0.01	179.62	<0.01
6.	I am afraid that people will find fault with me.	0.91	0.01	180.97	<0.01
8.	When I am talking to someone, I worry about what they may be thinking about me.	0.87	0.01	137.68	<0.01
9.	I am usually worried about what kind of impression I make.	0.87	0.01	135.07	<0.01
11.	Sometimes I think I am too concerned with what other people think of me.	0.86	0.01	124.38	<0.01
12.	I often worry that I will say or do the wrong things.	0.88	0.01	146.39	<0.01
Reverse-worded factor					<0.01
2.	I am unconcerned even if I know people are forming an unfavorable impression of me.	0.66	0.02	39.39	<0.01
4.	I rarely worry about what kind of impression I am making on someone.	0.67	0.02	39.57	<0.01
7.	Other people's opinions of me do not bother me.	0.80	0.01	58.51	<0.01
10.	If I know someone is judging me, it has little effect on me.	0.80	0.01	58.75	<0.01

CCFA categorical confirmatory factor analysis. *BFNE* brief fear of negative evaluation scale

J Psychopathol Behav Assess. Author manuscript; available in PMC 2016 September 16.

Table 2

Graded Model Item Parameters for Gender DIF Analysis

Item	*a*	*SE*	b_1	*SE*	b_2	*SE*	b_3	*SE*	b_4	*SE*
Graded Model Item Parameter Estimates for Women										
1	2.69	0.11	−1.08	0.05	−0.14	0.04	0.71	0.04	1.60	0.07
3	2.92	0.11	−0.88	0.05	0.09	0.04	0.79	0.04	1.78	0.07
5	3.78	0.16	−0.76	0.04	0.06	0.04	0.81	0.04	1.65	0.07
6	3.85	0.16	−0.72	0.04	0.10	0.04	0.81	0.04	1.69	0.07
8	3.31	0.14	−0.88	0.05	0.09	0.04	0.79	0.04	1.63	0.07
9	3.31	0.14	−1.08	0.05	−0.14	0.04	0.60	0.04	1.64	0.07
11	2.85	0.11	−0.89	0.04	−0.13	0.04	0.54	0.04	1.34	0.05
12	3.48	0.15	−0.85	0.04	0.03	0.04	0.71	0.04	1.57	0.06
Graded Model Item Parameter Estimates for Men										
1	2.42	0.16	−1.01	0.09	−0.08	0.06	0.74	0.07	1.80	0.11
3	2.92	0.11	−0.88	0.05	0.09	0.04	0.79	0.04	1.78	0.07
5	3.80	0.25	−0.80	0.07	0.05	0.05	0.78	0.06	1.59	0.08
6	3.82	0.26	−0.76	0.07	0.14	0.05	0.79	0.06	1.68	0.10
8	2.92	0.19	−1.05	0.07	0.01	0.05	0.73	0.06	1.69	0.10
9	2.88	0.19	−1.20	0.09	−0.18	0.05	0.62	0.06	1.72	0.10
11	2.85	0.11	−0.89	0.04	−0.13	0.04	0.54	0.04	1.34	0.05
12	3.11	0.20	−0.95	0.07	−0.07	0.05	0.70	0.06	1.49	0.09

The *a* parameter refers to the discrimination and the b_1,…, b_4 refer to the four threshold parameters

Table 3

DIF Statistics for Gender

Item	Total X^2	*df*	*BH-p*	X^2_a	*df*	*BH-p*	X^2_b	*df*	*BH-p*
1	7.60	5	0.663	2.00	1	0.237	5.60	4	0.989
5	0.90	5	0.972	0.00	1	0.966	0.90	4	0.989
6	1.90	5	0.972	0.00	1	0.966	1.90	4	0.989
8	5.80	5	0.663	2.80	1	0.237	3.00	4	0.989
9	3.70	5	0.898	3.40	1	0.237	0.30	4	0.989
12	6.80	5	0.663	2.20	1	0.237	4.60	4	0.989

The total X^2 refers to the omnibus test for DIF, the X^2_a refers to the test for non-uniform DIF, and the X^2_b refers to the test for uniform DIF

J Psychopathol Behav Assess. Author manuscript; available in PMC 2016 September 16.

Table 4

Graded Model Item Parameters for Ethnic DIF Analysis

Item	*a*	*SE*	b_1	*SE*	b_2	*SE*	b_3	*SE*	b_4	*SE*
Graded Model Item Parameter Estimates for Asian Participants										
1	2.58	0.09	−1.05	0.05	−0.11	0.03	0.73	0.04	1.68	0.06
3	2.95	0.10	−0.86	0.04	0.10	0.03	0.80	0.04	1.77	0.06
5	4.31	0.59	−0.70	0.13	0.03	0.08	0.73	0.07	1.60	0.16
6	3.67	0.54	−0.75	0.15	0.14	0.08	0.90	0.10	1.67	0.17
8	3.57	0.44	−0.65	0.12	0.22	0.08	0.86	0.11	1.87	0.20
9	4.28	0.63	−0.79	0.12	−0.00	0.08	0.68	0.09	1.60	0.16
11	3.54	0.43	−0.63	0.12	0.05	0.08	0.57	0.09	1.23	0.12
12	3.69	0.52	−0.85	0.15	0.12	0.08	0.71	0.09	1.56	0.18
Graded Model Item Parameter Estimates for Black Participants										
1	2.58	0.09	−1.05	0.05	−0.11	0.03	0.73	0.04	1.68	0.06
3	2.95	0.10	−0.86	0.04	0.10	0.03	0.80	0.04	1.77	0.06
5	3.87	0.68	−0.57	0.12	0.22	0.11	0.98	0.11	1.81	0.17
6	3.48	0.54	−0.62	0.13	0.27	0.10	0.89	0.11	1.92	0.21
8	2.55	0.30	−1.20	0.17	0.05	0.11	0.80	0.13	1.61	0.19
9	2.04	0.25	−1.55	0.20	−0.39	0.14	0.60	0.15	1.78	0.27
11	2.45	0.36	−0.86	0.16	−0.13	0.13	0.50	0.13	1.22	0.16
12	2.71	0.30	−0.76	0.12	−0.09	0.11	0.75	0.15	1.71	0.18
Graded Model Item Parameter Estimates for White Participants										
1	2.58	0.09	−1.05	0.05	−0.11	0.03	0.73	0.04	1.68	0.06
3	2.95	0.10	−0.86	0.04	0.10	0.03	0.80	0.04	1.77	0.06
5	3.82	0.15	−0.77	0.04	0.07	0.03	0.80	0.04	1.60	0.05
6	3.96	0.15	−0.72	0.04	0.11	0.03	0.79	0.04	1.65	0.06
8	3.27	0.13	−0.91	0.04	0.06	0.03	0.75	0.04	1.62	0.06
9	3.28	0.12	−1.09	0.04	−0.14	0.03	0.60	0.04	1.65	0.06
11	2.88	0.11	−0.89	0.04	−0.13	0.03	0.55	0.04	1.36	0.05
12	3.43	0.13	−0.87	0.04	0.01	0.03	0.71	0.04	1.51	0.05

The *a* parameter refers to the discrimination, the b_1, …, b_4 refer to the four threshold parameters

J Psychopathol Behav Assess. Author manuscript; available in PMC 2016 September 16.

Table 5

DIF Statistics for Ethnicity

Stem (Item)	Contrast	Total X^2	*df*	*BH-p*	X^2_a	*df*	*BH-p*	X^2_b	*df*	*BH-p*
Afraid not approve (5)	WH vs. BL	5.30	5	0.669	0.00	1	0.945	5.30	4	0.630
	WH vs. AS	1.40	5	0.927	0.70	1	0.626	0.70	4	0.950
Worry find fault (6)	WH vs. BL	4.50	5	0.696	0.70	1	0.626	3.70	4	0.847
	WH vs. AS	1.60	5	0.927	0.30	1	0.682	1.40	4	0.926
Worry thinking about me (8)	WH vs. BL	7.70	5	0.659	4.90	1	0.110	2.80	4	0.847
	WH vs. AS	5.90	5	0.659	0.40	1	0.682	5.50	4	0.630
Worried impression (9[a])	**WH vs. BL**	**30.80**	**5**	**0.001**	**19.50**	**1**	**0.001**	**11.30**	4	0.276
	WH vs. AS	4.20	5	0.696	2.40	1	0.335	1.80	4	0.926
Concerned others think (11)	WH vs. BL	6.60	5	0.659	1.30	1	0.510	5.30	4	0.630
	WH vs. AS	5.30	5	0.659	2.20	1	0.335	3.10	4	0.847
Worry do wrong things (12)	WH vs. BL	13.70	5	0.106	4.90	1	0.110	8.80	4	0.392
	WH vs. AS	2.80	5	0.878	0.20	1	0.682	2.60	4	0.847

The Stem refers to several words describing the item on BFNE-S (see Table 1 for full description). The total X^2 refers to the omnibus test for DIF, the X^2_a refers to the test for non-uniform DIF, and the X^2_b refers to the test for uniform DIF

WH white participants, *BL* black participants, and *AS* Asian participants

[a]Item nine reads “I am usually worried about what kind of impression I make”