Contents lists available at ScienceDirect

Personality and Individual Differences

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

The PHQ-9 assesses depression similarly in men and women from the general population

Michel A. Thibodeau, Gordon J.G. Asmundson *

Department of Psychology, University of Regina, Canada

ARTICLE INFO

Article history:
Received 1 June 2013
Received in revised form 22 August 2013
Accepted 30 August 2013
Available online 23 September 2013

Keywords:
PHQ-9
Differential item functioning
Depression
Psychometrics
Psychological assessment
Self-report questionnaires

ABSTRACT

Gender-based differential item functioning occurs when men and women respond differently to an item despite being similar on the trait assessed by that item. The Patient Health Questionnaire-9 (PHQ-9) is a prominent screening tool for depression. Researchers exploring whether the PHQ-9 exhibits gender-based differential item functioning have used only specialized samples (e.g., individuals with cancer or vision loss). We explored gender bias in the PHQ-9 by means of differential item functioning analyses in a population-based sample.

We made use of the National Health and Nutrition Examination Surveys (NHANES, 2008), a population-based sample of the USA including 5995 participants. Differential item functioning was assessed using the Mantel-Haenszel chi-square test and by comparing item characteristic curves between men and women.

All items exhibited negligible differential item functioning as demonstrated by the Mantel-Haenszel test, with absolute standardized mean differences ranging from 0.00 to 0.06. Item characteristic curves were similar between genders for all but one item. Item 5 (i.e., changes in appetite) exhibited very minor non-uniform differential item functioning, wherein extremely depressed women endorsed higher response options on this item compared to equally depressed men.

Researchers can use the PHQ-9 without concern of gender biases, particularly in epidemiological research.

1. Introduction

Differential item functioning is a form of measurement bias that occurs when an item's response properties vary across groups. In other words, differential item functioning leads to individuals from different groups responding differently to an item despite being similar on the trait assessed by that item. For example, women could respond differently to an item on an IQ test compared to men who share similar levels of intelligence. This bias is referred to as gender-based differential item functioning, which has traditionally received attention in the context of education and standardized achievement testing (Gierl, Khaliq, & Boughton, 1999; Penfield & Lam, 2000).

Gender-based differential item functioning can also impact other areas, such as epidemiological research and clinical endeavors (Teresi & Fleishman, 2007). Consider a self-report questionnaire that inflates symptoms in women due to biases in measurement. Such a bias could lead to hundreds or thousands of women being misclassified with regards to diagnostic status when using pre-designated cut-off scores in epidemiological studies. The prevalence or severity of a condition could thus be skewed on a wide scale, perhaps launching misguided efforts for future research (e.g., why is disorder X more prevalent in women?). Moreover, clinicians using measures that exhibit gender-based differential item functioning could conclude that certain patients have more (or less) severe symptoms than they actually do, leading to inaccurate screening and perhaps exclusion from potential services.

The Patient Health Questionnaire-9 (PHQ-9; Kroenke, Spitzer, & Williams, 2001) is a measure of depression that is prominent in practice and that is gaining popularity in epidemiological projects. Researchers have tested for gender-based differential item functioning in the PHQ-9 using a variety of samples, including individuals with HIV (Crane et al., 2010), with vision loss (Lamoureux et al., 2009), with traumatic brain injuries (Cook et al., 2011), with spinal cord injuries (Graves & Bombardier, 2008), with cancer (Smith et al., 2009), and in samples of individuals participating in heart surgery (Kendel et al., 2010) or who have been referred for mental health service (Cameron, Crawford, Lawton, & Reid, 2013). Findings have been mixed, with some studies suggesting that items on the PHQ-9 exhibit gender-based differential item

* Corresponding author. Address: Department of Psychology, University of Regina, Regina, Saskatchewan S4S 0A2, Canada. Tel.: +1 (306) 347 2415; fax: +1 (306) 337 3275.

E-mail address: gordon.asmundson@uregina.ca (G.J.G. Asmundson).

0191-8869/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.paid.2013.08.039

functioning (Crane et al., 2010; Graves & Bombardier, 2008; Kendel et al., 2010), and others finding no such characteristic (Cameron et al., 2013; Cook et al., 2011; Lamoureux et al., 2009; Smith et al., 2009). The use of diverse samples precludes conclusions regarding gender-based differential item functioning in populations outside of these specialized settings. Such conclusions could be useful to researchers and clinicians using the PHQ-9 outside of these areas.

Despite the recent thrust of research in this area, tests of gender-based differential item functioning in the PHQ-9 have yet to be applied to non-specialized samples or to samples representative of the general population. Researchers are frequently administering the PHQ-9 in large epidemiological surveys without knowing whether the measure assesses for depression in each gender without bias. Our aim was to test for gender-based differential item functioning in the PHQ-9 using a large epidemiological sample representative of the USA population. We tested for differential item functioning using the Mantel-Haenszel chi-square test and by comparing item characteristic curves (a feature of item response theory).

2. Material and methods

2.1. Participants

Data were from the National Health and Nutrition Examination Survey (NHANES; Centers for Disease Control & Prevention, 2008). The NHANES surveys were designed to assess the health of Americans by combining interviews and physical examinations. NHANES surveys have been conducted annually by the National Center for Health Statistics since 1999. The surveys make use of multistage sampling to allow inferences representative of noninstitutionalized adults and children in the USA. A four-stage sampling design is used, which focuses on US counties, census blocks, households, and persons. Persons 60 and older, African Americans, and Hispanics are oversampled. We used the 2008 version of the NHANES, which includes 5995 persons 18 and older. The survey included 3041 women (51%) and 2954 men (49%). Approximately 46% of participants reported their ethnicity as Non-Hispanic white, 21% reported being Non-Hispanic Black, 17% reported being Mexican American, 11% reported being other Hispanic, and 4% reported other ethnicities. Mean age of the sample was 49 (SD = 18.71). The survey was administered by trained interviewers in the households of participants. The portion of the survey included in this study was completed by a computer-assisted personal interview system. Participants were presented with a notebook computer and an electronic pen. The notebook computer displayed questions (e.g., items of the PHQ-9) and participants were asked to select their response with the pen.

2.2. The PHQ-9

The PHQ-9 includes nine Likert scale items that assess symptoms of depression as delineated in the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (American Psychiatric Association, 2000). All items share the header "Over the last 2 weeks, how often have you been bothered by any of the following problems?" All items (outlined in Table 1) share the same response options ranging from 0 ("*Not at all*") to 3 ("*Nearly every day*"). Cronbach's alpha for the PHQ-9 was .84 in the current sample.

2.3. Analyses

Differential item functioning requires a pre-determined latent factor structure. The PHQ-9 theoretically represents a single factor reflecting depression symptoms, which has been supported by previous research (Cameron, Crawford, Lawton, & Reid, 2008; Huang, Chung, Kroenke, Delucchi, & Spitzer, 2006; Merz, Malcarne, Roesch, Riley, & Sadler, 2011). We made use of confirmatory factor analysis to confirm the unidimensionality of the PHQ-9 prior to subsequent analyses. Confirmatory factor analysis was conducted in AMOS. Model fit was deemed acceptable if meeting the following values: Comparative Fit Index (CFI) value greater than .90, and exceeds or approaches .95; Root Mean Square Error of Approximation (RMSEA) value must be less than .08 and ideal fit approaches or is less than .05; and Standardized Root Mean Square Residual (SRMR) value should approach or fall below .08 (Browne & Cudeck, 1992; Tabachnick & Fidell, 2012). The chi-square test was not used as an index of model fit since the test is almost always statistically significant when using large samples.

Gender-based differential item functioning in the PHQ-9 was first tested with the Mantel-Haenszel chi-square test using jMetrik 2.1.0 (created by Patrick Meyer, University of Virginia, www.itemanalysis.com). The Mantel-Haenszel chi-square test makes use of a contingency table that simultaneously delineates an individual's response to an item, group status of the individual (e.g., man or woman), and scores on the measure from each group (Hidalgo & Lopez-Pina, 2004). Values of statistical significance for the Mantel-Haenszel tests were reported in conjunction with effect sizes to prevent exaggerated interpretation of negligible differential item functioning (Monahan, McHorney, Stump, & Perkins, 2007). Mantel-Haenszel values were classified to reflect one of the three following levels of differential item functioning: negligible differential item functioning when the Mantel-Haenszel test was non-significant or if absolute value of the standardized mean difference was less 0.15; marginal differential item functioning when the Mantel-Haenszel test was statistically significant and the absolute value of the standardized mean difference was greater than 0.15 and less than 0.30; and, definite differential item functioning when the Mantel-Haenszel test was statistically significant and the absolute value of the standardized mean difference was greater than 0.30 (Monahan et al., 2007; Zwick & Thayer, 1996).

Gender-based differential item functioning was subsequently explored by comparing item characteristic curves of men and women. Item characteristic curves are a facet of item response theory and plot which response option on an item is most likely to be endorsed by individuals along the continuum of a latent trait. For example, item characteristic curves may demonstrate that severely depressed men are most likely to endorse option "*3-Nearly every day*" on a question assessing loss of interest, while similarly depressed women may be more likely to endorse option "*2-More than half the day*". The distance between the curves thus serves as an index of differential item functioning. Item characteristic curves allow the examination of non-uniform differential item functioning, which is demonstrated by item characteristic curves that are separate only on part of the continuum (e.g., men and women respond differently to an item only when they are severely depressed). The latent trait of interest in this case is conceptualized as depression severity, which was estimated using all items of the PHQ-9. Item characteristic curves were rendered using jMetrik 2.1.0 and were smoothed using a Gaussian kernel.

3. Results

3.1. Descriptive statistics and dimensionality

The unidimensionality of the PHQ-9 was supported by the confirmatory factor analysis ($\chi^2 = 925.25$, df = 27, CFI = .94, RMSEA = 0.08 [lower limit 90% confidence interval = .07, upper limit 90% confidence interval = .08], SRMR = 0.04). Beta weights

Table 1
Differential item functioning on the PHQ-9 as demonstrated by the Mantel-Haenszel chi-square test.

PHQ-9 items		χ^2	Effect size	LL,UL 95% CI
1	Little interest or pleasure in doing things	6.20*	0.03	(0.01, 0.05)
2	Feeling down, depressed, or hopeless	1.26	−0.01	(−0.03, 0.01)
3	Trouble falling or staying asleep, or sleeping too much	0	0	(−0.03, 0.03)
4	Feeling tired or having little energy	1.42	−0.02	(−0.04, 0.01)
5	Poor appetite or overeating	23.47**	−0.06	(−0.08, −0.03)
6	Feeling bad about yourself—or that you are a failure or have let yourself or your family down	0	0	(−0.02, 0.02)
7	Trouble concentrating on things, such as reading the newspaper or watching television	8.94**	0.03	(0.01, 0.05)
8	Moving or speaking so slowly that other people could have noticed. Or the opposite—being so fidgety or restless that you have been moving around a lot more than usual	2.15	0.01	(0.00, 0.03)
9	Thoughts that you would be better off dead or of hurting yourself in some way	7.91**	0.02	(0.01, 0.03)

Note: LL, lower limit; UL, upper limit; CI, confidence interval.
* $p < .01$.
** $p < .001$.

for items ranged from .48 (item 9) to .77 (item 2) and were all statistically significant ($p < .001$). The PHQ-9 was thus considered as comprising one factor of nine items for subsequent analyses. Women scored significantly greater than men with respect to PHQ-9 total scores (3.93, SD = 4.71 vs 2.72, SD = 3.89, respectively; $t = 10.34$, $p < .001$). Approximately 10% of the sample (187 men, 343 women) scored 10 or above, which has been denoted as the suggested cut-off for depression using the PHQ-9 (Kroenke et al., 2001).

3.2. Mantel-Haenszel test results

Results of the Mantel-Haenszel chi-square test and estimated effect sizes of differential item functioning are reported in Table 1. A total of 4 Mantel-Haenszel chi-square tests reached statistical significance (items 1, 5, 7, and 9); however, all effect sizes were very small and in the negligible range (absolute effect sizes from 0.00 to 0.06).

3.3. Item characteristic curves

Item characteristic curves are reported in Fig. 1 and were nearly identical for men and women. Women and men shared marginally dissimilar item characteristic curves for item 5 ("*Poor appetite or overeating*"), with distance between the curves increasing along the continuum of depression severity. Specifically, the curves are nearly identical for individuals up to 1 standard deviation above the mean in depression severity, and the lines separate progressively as levels of depression increase. Women scoring 1.5 standard deviations above the mean and greater with respect to depression severity are expected to select approximately half a response option greater (e.g., selecting response option 2) than men with similar levels of depression (e.g., selecting on average a hypothetical response option of 1.5).

4. Discussion

We used two indices of gender-based differential item functioning and a population-based sample to determine if the PHQ-9 assesses depression similarly in men and women. The results from both analyses were congruent. The Mantel-Haenszel chi-square test demonstrated that no item exhibited gender-based differential item functioning beyond a negligible range. Moreover, our interpretation of item characteristic curves did not highlight curves that contrasted significantly between men and women. Item 5, which assesses for changes in appetite, was an exception. The most depressed women were likely to score higher on the item (i.e., endorse changes in eating habits) compared to men who share similar levels of depression. Gender-based differential item functioning was not present in men and women who score up to one standard deviation over the mean (i.e., the bias is not uniform across symptoms of depression). Moreover, the differential item functioning is unlikely to have any noticeable impact. Indeed, depressed women seem to score at most half a point above men who share similar levels of depression.

Our use of item characteristic curves also highlighted features of the PHQ-9 that may warrant further research. The items did not differentiate between individuals with extremely low (i.e., below two standard deviations under the mean) to mean levels of depression. In other words, individuals who are not at all depressed are most likely to endorse response option "0" on each of the items, but so are individuals who are average in depression or perhaps even slightly above the mean. This lack of discrimination may not impact clinical endeavors, but could plausibly impact research exploring non-clinical depression. Moreover, items 8 (psychomotor agitation/retardation) and 9 (thoughts of self-harm) provided particularly little information. Individuals who endorsed extreme symptoms of depression were likely to only choose response option "1" or "2" on the Likert scale for item 8. Item 9 was even worse, with extremely depressed individuals selecting only option "1", with options "2" and "3" being of little use. Researchers may consider how to change the content of these items to make better use of all response options and to assess the full breadth of depression symptoms.

The current report adds to growing literature on the PHQ-9 and supports the utility of the measure. More research is needed to assess for gender-based differential item functioning in other instruments of depression. For example, our research group has recently identified substantial differential item functioning in an item of the popular Center for Epidemiologic Studies Depression Scale (Carleton et al., 2013; Radloff, 1977), which may impact epidemiological surveys, other forms of research, and clinical endeavors. Other instruments of depression likely share a similar characteristic. Moreover, other psychological instruments likely exhibit notable differential item functioning and research on these measures is needed to help ensure accurate and unbiased measurement of psychological constructs.

Our study is associated with a few limitations. Use of a sample that did not include institutionalized adults raises questions regarding the severity of symptoms in this study; however, many individuals with severe depression are likely included in the NHANES and we believe the span of depression symptoms in this study allows conclusions applicable to most clinical samples. Unfortunately, we were unable to use the statistical weights provided in the NHANES dataset due to the complexity of the analyses used in this study.

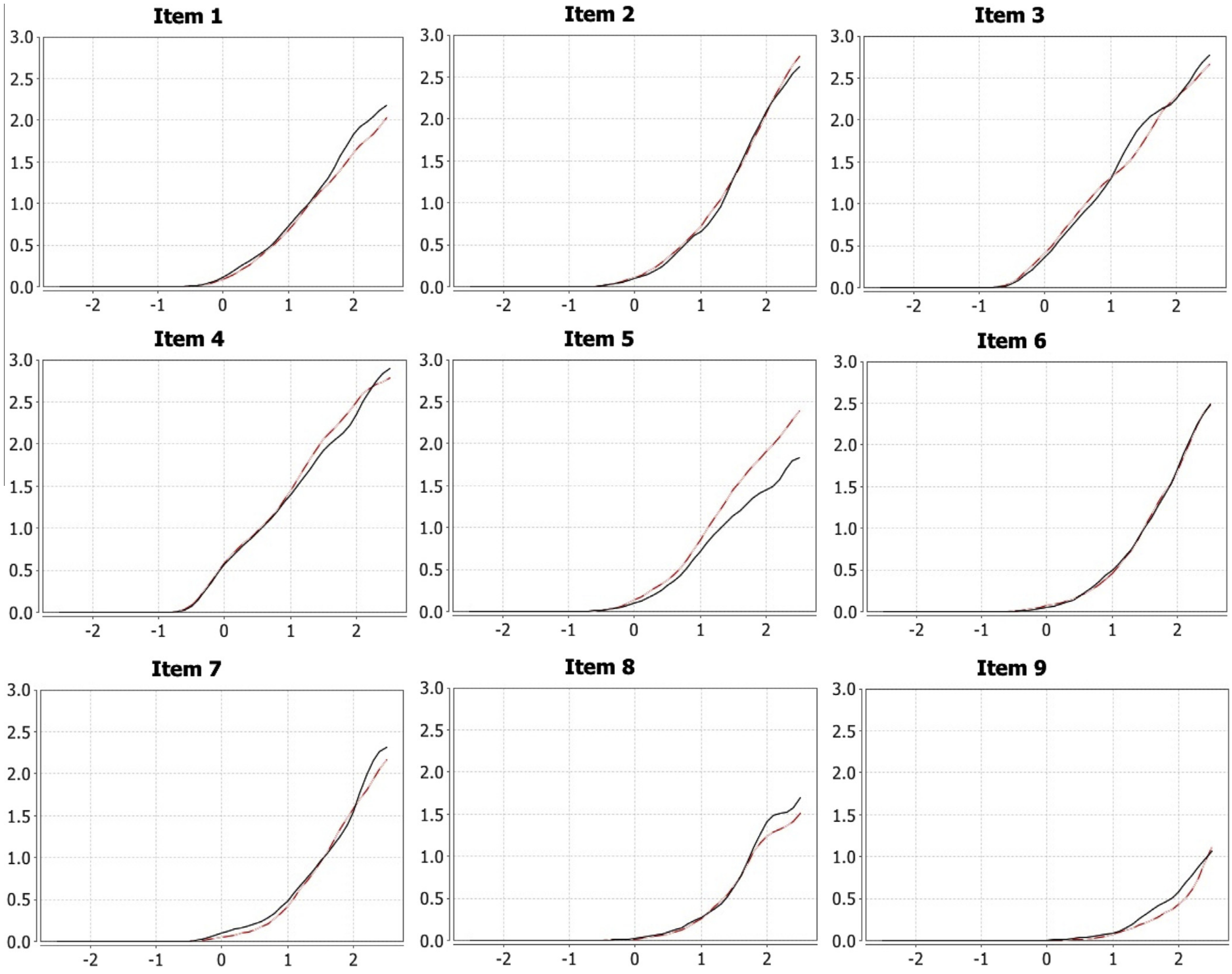

Fig. 1. Item characteristic curves of men and women for each item in the PHQ-9. The plots demonstrate the likely response option of men and women on the PHQ-9 Likert scale items (*Y*-axis) along the continuum of depression severity (*X*-axis). Scores on the *X*-axis represent standard normalized deviations. The solid lines represent men, the dotted lines represent women.

Notwithstanding, the stratified and heterogeneous nature of the sample allows relatively broad inferences to the general population.

Researchers can confidently use the PHQ-9 without worry of gender biases, particularly in the context of epidemiological research. Our use of two indices of gender-based differential item functioning and of a population-based sample highlights the robustness of the present findings. Researchers may consider exploring gender-based differential item functioning in other prominent self-report measures used in research and practice.

Acknowledgements

The authors declare no conflicts of interest. Michel A. Thibodeau is supported by a Canadian Institutes of Health Research doctoral award (FRN: 113434).

References

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

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*(2), 230–258.

Cameron, I. M., Crawford, J. R., Lawton, K., & Reid, I. C. (2008). Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. *British Journal of General Practice, 58*(546), 32–36. http://dx.doi.org/10.3399/bjgp08X263794.

Cameron, I. M., Crawford, J. R., Lawton, K., & Reid, I. C. (2013). Differential item functioning of the HADS and PHQ-9: An investigation of age, gender and educational background in a clinical UK primary care sample. *Journal of AffectiveDisorders*, 147(1-3), 262–268.

Carleton, R. N., Thibodeau, M. A., Teale, M. J. N., Welch, P. G., Abrams, M. P., Robinson, T., et al. (2013). The Center for Epidemiologic Studies Depression Scale: A review with a theoretical and empirical examination of item content and factor structure. *PLOS One*, e58067. http://dx.doi.org/10.1371/journal.pone.0058067.

Centers for Disease Control and Prevention (2008). *National Health and Nutrition Examination Survey Data (2007–2008)*. US Department of Health and Human Services.

Cook, K. F., Bombardier, C. H., Bamer, A. M., Choi, S. W., Kroenke, K., & Fann, J. R. (2011). Do somatic and cognitive symptoms of traumatic brain injury confound depression screening? *Archives of Physical Medicine and Rehabilitation, 92*(5), 818–823. http://dx.doi.org/10.1016/j.apmr.2010.12.008.

Crane, P. K., Gibbons, L. E., Willig, J. H., Mugavero, M. J., Lawrence, S. T., Schumacher, J. E., et al. (2010). Measuring depression levels in HIV-infected patients as part of routine clinical care using the nine-item Patient Health Questionnaire (PHQ-9). *AIDS Care, 22*(7), 874–885. http://dx.doi.org/10.1080/09540120903483034.

Gierl, M., Khaliq, S. N., & Boughton, K. (1999). *Gender differential item functioning in mathematics and science: Prevalence and policy implications*. Paper presented at the Annual Meeting of the Canadian Society for the Study of Education, Sherbrooke, Québec, Canada.

Graves, D. E., & Bombardier, C. H. (2008). Improving the efficiency of screening for major depression in people with spinal cord injury. *Journal of Spinal Cord Medicine, 31*(2), 177–184.

Hidalgo, M. D., & Lopez-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*(6), 903–915. http://dx.doi.org/10.1177/0013164403261769.

Huang, F. Y., Chung, H., Kroenke, K., Delucchi, K. L., & Spitzer, R. L. (2006). Using the Patient Health Questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. *Journal of General Internal Medicine, 21*(6), 547–552. http://dx.doi.org/10.1111/j.1525-1497.2006.00409.x.

Kendel, F., Wirtz, M., Dunkel, A., Lehmkuhl, E., Hetzer, R., & Regitz-Zagrosek, V. (2010). Screening for depression: Rasch analysis of the dimensional structure of the PHQ-9 and the HADS-D. *Journal of Affective Disorders, 122*(3), 241–246. http://dx.doi.org/10.1016/j.jad.2009.07.004.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine, 16*(9), 606–613.

Lamoureux, E. L., Tee, H. W., Pesudovs, K., Pallant, J. F., Keeffe, J. E., & Rees, G. (2009). Can clinicians use the PHQ-9 to assess depression in people with vision loss? *Optometry & Vision Science, 86*(2), 139–145. http://dx.doi.org/10.1097/OPX.0b013e318194eb47.

Merz, E. L., Malcarne, V. L., Roesch, S. C., Riley, N., & Sadler, G. R. (2011). A multigroup confirmatory factor analysis of the Patient Health Questionnaire-9 among English- and Spanish-speaking Latinas. *Cultural Diversity & Ethnic Minority Psychology, 17*(3), 309–316. http://dx.doi.org/10.1037/a0023883.

Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics, 32*(1), 92–109. http://dx.doi.org/10.3102/1076998606298035.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19*(3), 5–15. http://dx.doi.org/10.1111/j.1745-3992.2000.tb00033.x.

Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385–401.

Smith, A. B., Rush, R., Wright, P., Stark, D., Velikova, G., & Sharpe, M. (2009). Validation of an item bank for detecting and assessing psychological distress in cancer patients. *Psychooncology, 18*(2), 195–199. http://dx.doi.org/10.1002/pon.1423.

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Boston: Pearson Education.

Teresi, J. A., & Fleishman, J. A. (2007). Differential item functioning and health assessment. *Quality of Life Research, 16*(Suppl. 1), 33–42. http://dx.doi.org/10.1007/s11136-007-9184-6.

Zwick, R., & Thayer, D. T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics, 21*(3), 187–201.