Published in final edited form as:

Assessment. 2016 December ; 23(6): 758–768. doi:10.1177/1073191115599639.

Measurement Equivalence of the K6 Scale: The Effects of Race/Ethnicity and Language

Giyeon Kim¹, Jamie DeCoster², Ami N. Bryant¹, and Katy L. Ford¹

¹The University of Alabama, Tuscaloosa, AL, USA

²University of Virginia, Charlottesville, VA, USA

Abstract

This study examined the measurement equivalence of the K6 across diverse racial/ethnic and linguistic groups in the U.S. differential item functioning analyses using item response theory were conducted among 44,846 U.S. adults drawn from the California Health Interview Survey. Results show that four items (“nervous,” “restless,” “depressed,” and “everything an effort”) varied significantly across races/ethnicities and four items (“nervous,” “hopeless,” “restless,” and “depressed”) varied significantly across languages. In additional effect size analyses designed to separate effects of race/ethnicity from language, the structure of the White English group was substantially different from both the Hispanic/Latino English group and Hispanic/Latino Spanish group, whereas the Hispanic/Latino Spanish group was not different from the Hispanic/Latino English group. The findings suggest that there was evident measurement nonequivalence in the K6 among racially/ethnically and linguistically diverse adults and that the observed nonequivalence in the K6 appears to be driven by language rather than race/ethnicity.

Keywords

K6; measurement equivalence; differential item functioning (DIF); race/ethnicity; language

The K6 scale (Kessler et al., 2002) is a widely used screening instrument that was developed to screen for nonspecific psychological distress in the general population included in large health surveys such as the National Health Interview Survey (NHIS) and the National Household Survey on Drug Abuse (NHSDA). The K6 consists of six items embedded within a 10-item scale (i.e., K10; Kessler et al., 2002), and the K10 items were chosen from an original pool of 612 pilot items using item response theory (IRT) models. The resultant K6 items measures feelings of nervousness, hopelessness, restlessness, depression, worthlessness, and that everything is an effort, after deleting four questions (“tired out for no good reason,” “so nervous that nothing could calm you down,” “so restless that you could not sit still,” and “depressed”) from the K10. Developed for use in epidemiological studies,

Reprints and permissions: sagepub.com/journalsPermissions.nav

Corresponding Author: Giyeon Kim, Alabama Research Institute on Aging and Department of Psychology, The University of Alabama, Box 870315, Tuscaloosa, AL 35487-0315, USA., giyeon.kim@ua.edu.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

the K6 offers an alternative to lengthy diagnostic interviews, providing a measure of symptom severity and overall levels of distress, rather than a specific diagnosis (Kessler et al., 2002). The advantage of the six-item short form (i.e., K6) for use in large community epidemiological surveys is its brevity and maximized precision in the clinical range of the scale (Kessler et al., 2002). Based on its pilot studies and on its inclusion in the 1997 and 1998 NHIS, Kessler et al. (2002) reported excellent precision of this six-item scale (i.e., K6) in terms of the scale distribution and severity across major sociodemographic subsamples such as those based on age, sex, and education. With total scores ranging from 0 to 24, a standard cutoff score of 13 or higher on the K6 has been used to identify persons with nonspecific serious psychological distress (SPD; i.e., those with a high likelihood of having a diagnosable mental illness severe enough to cause functional limitations and to require treatment; Furukawa et al., 2003; Kessler et al., 2003; Kim, Bryant, & Parmelee, 2012; Pratt, 2009).

Originally developed in English, the K6 has been used in more than 30 countries and has been translated into several different languages including Arabic, Chinese, Italian, Japanese, and Spanish. Previous studies reported good utility of the translated versions of the K6 (e.g., Furukawa et al., 2008; Lee et al., 2012; Sakurai et al., 2011). For example, Lee et al. (2012) found that the Chinese version of the K6 had substantial concordance with face-to-face clinical interviews in Hong Kong, concluding that the Chinese version of the K6 is a valuable tool for screening serious mental illness. As another example, Furukawa et al. (2008) reported that the Japanese version of the K6 demonstrated screening performances equivalent to the original English version. However, some researchers raised concerns about the validity of translated versions of the K6 such as the Spanish version (McVeigh et al., 2006). For example, McVeigh et al. (2006) reported the high prevalence of psychological distress among people interviewed in Spanish and noted that given the K6 has not been validated in Spanish, bias introduced by use of the Spanish translated version of the K6 might have contributed to their findings. Thus, it is not yet clear whether the translated versions of the K6 are equivalent across different languages.

The K6 has also been used to screen serious mental illness in diverse racial, ethnic, and cultural groups (e.g., Andersen et al., 2011; Kim, Bryant, & Parmelee, 2012; Kim et al., 2010; Mitchell & Beals, 2011). In the initial stage of the scale development conducted by Kessler et al. (2002), they found that 38 of 53 items included had consistent severity parameters across sociodemographic subsamples including race/ethnicity. However, previous studies reported mixed findings on the performance of the K6 when it is applied to diverse cultural groups. Some previous research has reported the successful use of the K6 in diverse racial/ethnic groups (e.g., Mitchell & Beals, 2011). For example, Mitchell and Beals (2011) reported that the K6 was shown to be an appropriate screening and severity measure for mood disorders in two American Indian samples included in the American Indian Service Utilization, Psychiatric Epidemiology, and Risk and Protective Factors Project. The authors recommended inclusion of a measure such as the K6 as a complement to more traditional dichotomous diagnoses in both research and clinical settings.

Conversely, there are studies raising concerns about cross-cultural comparability of the K6 (e.g., Andersen et al., 2011; Kim et al., 2012). Andersen et al. (2011) examined the

psychometric properties of the K6 when screening for mood and anxiety disorders between Blacks (majorities) and others (minorities) in South Africa and found that the K6 was less successful in screening for depression and anxiety disorders among Blacks than among minorities. The authors concluded that the K6 is not as useful when screening for depression and/or anxiety disorders among general population of South Africa. Additionally, in a study using the K6 to assess SPD among diverse older adults in the United States, Kim et al. (2012) found significant racial/ethnic differences in the prevalence of past year SPD showing the American Indian/Alaska Native (AI/AN) group having the highest prevalence (18.23%) and the Asian group having the lowest prevalence (2.61%). However, the authors noted that findings should be interpreted with caution as the measurement equivalence of the K6 across diverse racial/ethnic groups has not yet been established (Kim et al., 2012).

Establishing measurement equivalence across diverse racial/ethnic and cultural groups is a fundamental issue in cross-cultural research. If measures do not function equivalently across different groups, cross-cultural comparisons using these measures may not be accurate and results from these comparisons would be misleading. Differential item functioning (DIF) occurs when individuals from different groups show unequal probabilities of endorsing specific items despite equal levels of the underlying construct that the item is intended to measure (Zumbo, 1999). This suggests that people from varying racial/ethnic, linguistic, or cultural groups may show different probabilities of answering items in the face of equivalent scores on psychological distress. It should be also noted that DIF is a necessary, but not sufficient, condition for item bias (Zumbo, 1999). Item bias occurs when differential responding is related to some characteristics of the measure that is not relevant to the measure's purpose (Zumbo, 1999). For example, Teresi, Ramirez, Lai, and Silver (2008) reviewed DIF studies on patient-reported outcome measures such as depression, quality of life, and general health and reported that some potential sources of DIF were age, gender, education, race/ethnicity, language, and chronic conditions.

There are several important reasons to study measurement equivalence of the K6 across racially/ethnically and linguistically diverse groups. The K6 has been included as a screening measure in many national surveys (e.g., NHIS and NHSDA) and results from these national surveys are often presented in many federal reports (e.g., National Healthcare Disparities Report and National Healthcare Quality Report). Such reports have the potential to influence important mental health policy decisions. Similarly, the concept of SPD that the K6 assesses is particularly useful in large-scale surveys as well as in the context of public policy analysis. In addition, although we tend to assume that the translation of the K6 is done adequately, variability in the quality of translation could be a more fundamental reason for the nonequivalent measures, which may be a topic that is worth investigating.

Given the inconsistent findings concerning the cross-racial/ethnic comparability of the K6 and the lack of knowledge concerning the cross-linguistic comparability of the K6 in the current literature, the present study examined the measurement equivalence of the K6 across diverse racial/ethnic and linguistic groups in the United States. While language can be viewed as one of many components of ethnicity, the two are distinct concepts (Zagefka, 2009). These concepts are defined in the methods section; however, an example helps illustrate the distinction. One can speak only English and identify as Hispanic/Latino race/

ethnicity or one can identify as Black and speak only Spanish. In addition, being a racial/ethnic minority does not equate to interviewing in non-English languages. Therefore, we sought to separate the effects of race/ethnicity and language on measurement equivalence of the K6. Due to the exploratory nature of this study, no specific hypotheses were proposed. Given the wide use of the K6 scale in many national epidemiology surveys, results from this study will provide useful information for future cross-cultural studies using the K6 scale.

Method

Sample

Participant data were drawn from the 2009 California Health Interview Survey (CHIS) that was conducted between September 2009 and April 2010 by the UCLA Center for Health Policy Research in collaboration with the California Department of Health Services and the Public Health Institute. The CHIS is a biannual telephone survey conducted in the state of California. The CHIS examines a diverse sample by relying on a supplemented surname list sample and by conducting the survey in six different languages (English, Spanish, Korean, Cantonese, Mandarin, and Vietnamese) to capture language needs of diverse groups in California. For translation from English into other languages, the CHIS used the refereed multiple forward translations method, which involves an outside referee to judge the quality of each of the forward translations by professionally credentialed bilingual translators (Ponce et al., 2004). Additional information about the CHIS data collection and translation procedures is available on the CHIS webpage (CHIS, n.d.; Ponce et al., 2004). Use of the publicly available CHIS data set has been preapproved by the University of Alabama Institutional Review Board.

All 2009 CHIS participants ($n = 47,614$) were included for the initial overall analysis. The racial/ethnic composition of the entire CHIS sample was 66.7% non-Hispanic White ($n = 31,769$), 4.1% African American/Black ($n = 1,950$), 12.1% Hispanic/Latino ($n = 5,753$), 10.2% Asian ($n = 4,874$), 1.1% AI/AN ($n = 500$), 0.2% Pacific Islander ($n = 75$), and 5.7% Other ($n = 2,693$). Participants who did not fall into one of the major five racial/ethnic groups being analyzed for an individual IRT (non-Hispanic White, African American/Black, Hispanic/Latino, Asian or AI/AN) were excluded. In these five racial/ethnic groups, six languages (English, Spanish, Korean, Cantonese, Mandarin, and Vietnamese) were used for interviews. A total of 44,846 adults aged 18 years and older were used in our DIF analyses.

Measures

The K6 Scale—The K6 scale (Kessler et al., 2002) consists of six items to assess psychological distress. Respondents were asked to report, “During the past 30 days, about how often did you feel nervous (Item 1), hopeless (Item 2), restless or fidgety (Item 3), so depressed that nothing could cheer you up (Item 4), that everything was an effort (Item 5), and worthless (Item 6)?” Responses to each question ranged from 0 (*none of the time*) to 4 (*all of the time*). The total score ranged between 0 and 24, with higher scores indicating higher levels of psychological distress. A score of 13 or higher indicates SPD (Furukawa et al., 2003; Kessler et al., 2003). As mentioned earlier, the K6 has been translated into several different languages including Chinese, Italian, Japanese, and Spanish. The internal

consistency for each racial/ ethnic group included in the study was found to be satisfactory: $\alpha = .80$ for non-Hispanic Whites, $\alpha = .82$ for African Americans/Blacks, $\alpha = .81$ for Asians, $\alpha = .85$ for AI/ANs, and $\alpha = .82$ for Hispanics/Latinos. The internal consistency for each language group included in the study was found to be satisfactory as well: $\alpha = .80$ for English, $\alpha = .84$ for Spanish, $\alpha = .75$ for Vietnamese, $\alpha = .85$ for Korean, $\alpha = .90$ for Cantonese, and $\alpha = .85$ for Mandarin.

Race/Ethnicity—Defined by the federal Office of Management and Budget and the U.S. Census Bureau, race and ethnicity in the United States are self-identified categories of respondents' origins. Racial/ethnic categories reflect a social definition recognized in country, rather than defining race/ethnicity biologically, anthropologically, or genetically (Humes, Jones, & Ramirez, 2011). The CHIS provides self-reported racial/ethnic categories, which include non-Hispanic White, African American/Black, Hispanic/Latino, AI/ AN, and Asian.

Language—In the present study, the language variable was measured by the language in which the CHIS interview was conducted. Language categories included in the survey were English, Spanish, Korean, Cantonese, Mandarin, and Vietnamese.

Data Analysis

The primary aim of this study was to examine whether the items of the K6 scale function differently across different races/ethnicities or languages. DIF of a scale across groups can be examined either using confirmatory factor analysis (CFA) or IRT. Stark, Chernyshenko, and Drasgow (2006) compared the performance of these two methods using Monte Carlo simulations. Their results support the use of IRT analyses with polytomous data when large samples with a single latent factor are used. With small differences, IRT tended to have slightly less power than CFA, but it also had notably lower Type I error rates. With large differences, the power was approximately equal, but the IRT still had lower error rates. We therefore decided to base our own analyses on IRT.

In order to test unidimensionality, which is important to the validity of IRT DIF analyses, we performed a CFA assuming that the K6 had a single factor with no correlations among the residuals for the individual items. This analysis found that the model had acceptable fit (root mean square error of approximation = .08, standardized root mean square residual = .04, comparative fit index = .94).

When using IRT to examine DIF, researchers must choose to use either a constrained baseline model or a free baseline model. In the constrained baseline model, researchers hold the structure of all of the items to be fixed between groups and then determine the model fit. Then they free one of the items and allow it to vary between groups. The extent to which this improves model fit is a measure of how much that item varies between groups. In the free baseline method, researchers assess the fit of a model where the structure of one item is constrained to be equal across all of the groups. Researchers then assess the fit of a model where a reference item and a second item are constrained to be equal across all of the groups. The difference between the two fit statistics is then taken as a measure of how much the second item varies between groups. Stark et al. (2006) showed that the free baseline

model is somewhat less likely to lead to Type I errors. However, the free baseline method does have a problem in that it does not provide a test of the reference item. As a resolution, we decided to initially use a constrained model to explore differences between groups, in each case identifying one item that does not vary between groups. We then used that item as the referent for a set of analyses using a free baseline. Given that the problem with the constrained baseline appears to be that it is too likely to identify group differences, we should be able to be confident in any findings suggesting that an item does not vary between groups.

DIF analyses in IRT were performed by comparing the fit of a full model (which allows more freedom for the structure to vary between groups) with a reduced model (which makes more assumptions about the equivalence of the structure across groups). The fit statistic we used is G^2 (i.e., $-2 \log$ likelihood), which provides a measure of the fit of the overall model (Stark et al., 2006). It is similar to the χ^2 difference test, but the G^2 provides a better approximation of the underlying distribution (Harremoës & Tusnády, 2012). The difference between the G^2 statistics of two nested models follows a χ^2 distribution, which can be used to test whether freeing certain constraints leads to an overall improvement of model fit (e.g., Bolt, 2002; Stark et al., 2006). In a constrained baseline model, a significant χ^2 difference test indicates that the structure of the freed item varies significantly between groups. In a free baseline model, a significant χ^2 difference test indicates that the constrained item varies significantly between groups.

IRT analyses were performed using IRTPRO version 2.1 (Cai, Thissen, & du Toit, 2011). We used a graded response model (Samejima, 1969, 1997) to estimate the response functions because the items took on values from 0 to 4. We used the Bock and Aitkin (1981) variant of the expectation maximization method to estimate item parameters, which is commonly used for unidimensional IRT analyses.

After testing whether there were significant differences between groups, we examined effect size measures of DIF to determine which groups were different from each other. Specifically, we calculated the Signed Test Difference in the Sample (STDS) and the Expected Test Score Standardized Difference (ETSSD) for each comparison, with the STDS representing the expected difference between the two groups in terms of the sum score for the scale, and the ETSSD providing a similar measure rescaled so that it is on the same metric as Cohen's d (Meade, 2010).

Results

Sample Description

As summarized in Table 1, all background characteristics of the CHIS sample varied significantly across the five racial/ ethnic groups (all $ps < .001$). The majority of the total sample included in the DIF analyses was interviewed in English (89.0%), followed by Spanish (6.9%), Vietnamese (2.0%), Korean (1.4%), Mandarin (0.4%), and Cantonese (0.4%). Among Hispanics/Latinos, over half were interviewed in Spanish (52.4%). While over half of Asians were interviewed in English (62.0%), other Asian languages were also

used for the CHIS interviews: Vietnamese (18.1%), Korean (12.6%), Mandarin (3.8%), and Cantonese (3.4%).

The mean age of the sample was 56.12 years ($SD = 17.22$, range: 18–85), with Hispanics/Latinos being the youngest ($M = 44.27$, $SD = 16.28$) and Whites ($M = 59.24$, $SD = 16.31$) being the oldest. More than half were female and married for all five racial/ethnic groups, although these characteristics did vary significantly between racial/ethnic differences. More females were included in the African American/Black sample (63.9%), whereas relatively fewer females were included in the Asian sample (56.2%). Asians were most likely to be married (65.9%), whereas African Americans/Blacks were least likely (33.8%). Over half of Whites, African Americans/Blacks, Asians, and AI/ANs had some college or higher education, whereas only one third of Hispanics/Latinos had college or higher education. Whites had the highest level of annual household income and Hispanics/Latinos had the lowest income level. The AI/AN group showed higher K6 scores and higher percentages of SPD, whereas the Asian and White groups had lower K6 scores and lower percentages of SPD.

Differential Item Functioning Analyses: The Effect of Race/Ethnicity

Constrained Baseline Analyses—Our initial constrained baseline IRT analyses indicated that all of the K6 items varied significantly across races/ethnicities ($p < .05$) except for Item 6 (“worthless”; $\chi^2[20] = 18.57$, $p = .55$). We therefore chose to use Item 6 (“worthless”) as the referent in our free baseline IRT analyses. It should be noted that Item 6 (“worthless”) is the only item that is fundamentally social in nature.

Free Baseline Analyses—Table 2 reports the Akaike information criterion (AIC), $-2 \log$ likelihood, and G^2 statistics for the free baseline IRT DIF analyses. The baseline model only constrains Item 6 (“worthless”) to be consistent across racial/ethnic groups. Each of the other models constrains Item 6 (“worthless”) as well as one other item to be consistent across racial/ethnic groups. The tests reported in the last three columns (i.e., G^2 difference from baseline, degrees of freedom difference from baseline, and chi-square p) in Table 2 determine whether constraining the second item leads to significantly worse fit relative to the baseline model. Results from DIF analyses show that Items 1, 3, 4, and 5 (“nervous,” “restless,” “depressed,” and “everything an effort,” respectively) vary significantly between races/ethnicities, but Item 2 (“hopeless”) does not. Although our free baseline analyses do not provide us with a test of Item 6 (“worthless”), based on the constrained baseline analyses, we can conclude that it does not vary by race/ethnicity.

In order to better elucidate the differences across racial/ethnic groups, we examined the effect sizes to determine exactly which racial/ethnic groups had structures that were significantly different from the structure for Whites. The estimates of these effect sizes for the comparison of the structures for each race with the structure found in Whites are presented in Table 3. Results indicated that the structure for Whites is minimally different from those for African Americans/Blacks and AI/ANs, but substantially different from that for Hispanics/Latinos and Asians. The total information curves for Whites, Hispanics/Latinos, and Asians (not presented) show that for Whites, the scale shows optimal

discrimination for individuals with high latent scores, whereas for Hispanics/ Latinos and Asians, the scale shows optimal discrimination for individuals with medium latent scores.

Differential Item Functioning Analyses: The Effect of Language

Constrained Baseline Analyses—Our initial constrained baseline IRT analyses indicated that all of the items varied significantly across languages ($p < .05$) except for Item 6 (“worthless”; $\chi^2[25] = 30.95, p = .19$). Therefore, Item 6 (“worthless”) was used as the referent in our free baseline IRT analyses.

Free Baseline Analyses—Table 4 reports the AIC, $-2 \log$ likelihood, and G^2 statistics for the free baseline IRT DIF analyses for the effect of language. The baseline model only constrains Item 6 (“worthless”) to be consistent across languages. Each of the other models constrains Item 6 (“worthless”) as well as one other item to be consistent across languages. The tests reported in the last three columns in Table 3 determine whether constraining the second item leads to significantly worse fit relative to the baseline model. Results from DIF analyses show that Items 1, 2, 3, and 4 (“nervous,” “hopeless,” “restless,” “depressed,” respectively) vary significantly across languages, but Item 5 (“everything an effort”) does not. Our free baseline analyses do not provide us with a test of Item 6 (“worthless”), but we can conclude that it does not vary between languages based on the constrained baseline analyses.

In order to better understand the differences across languages, we examined the effect sizes to determine which languages had structures that were different from the structure found in the English version. This was accomplished by calculating the STDS and the ETSSD for each comparison, which are presented in Table 5. Results indicated that the structure for English has small differences from the structures for Vietnamese, Korean, Mandarin, and Cantonese versions, but large differences from the Spanish version.

The total information curves for the English and Spanish versions are presented in the Supplement Figures (available online at <http://asm.sagepub.com/content/by/supplemental-data>). Results indicated that the English version of the K6 scale shows optimal discrimination for individuals with high latent scores, whereas the Spanish version shows optimal discrimination for individuals with medium latent scores.

Differential Item Functioning Analyses: Separating Effects of Hispanic/Latino Ethnicity From Spanish Language

The prior analyses indicated that there were significant differences in the structure of the K6 between Whites and Hispanics/Latinos as well as between the English-language and Spanish-language versions. However, given that being Hispanic/Latino is likely related to using the Spanish version of the scale, it is worth examining these effects more closely before drawing any firm conclusions. Table 6 presents a cross-tabulation representing the relation between race/ethnicity and language (after limiting our consideration to those involved in these comparisons). This clearly shows that those taking the English version were more likely to be White, whereas those taking the Spanish version were more likely to be Hispanic/Latino (ϕ coefficient = .69, $p < .001$).

To obtain a better understanding of the effects of Hispanic/Latino ethnicity and Spanish language, we performed an additional set of DIF analyses examining differences across the different combinations of race/ethnicity (White vs. Hispanic/Latino) and language (English vs. Spanish). However, the small number of White individuals who took the Spanish version of the K6 ($n = 52$) prevented us from being able to explore the structure of the scale within this particular combination. We therefore performed DIF analyses examining difference across the remaining three combinations: (a) Whites who took the English version, (b) Hispanics/Latinos who took the English version, and (c) Hispanics/Latinos who took the Spanish version. Given our prior results, we expected to observe differences between the structures of the K6 for Whites who took the English version and Hispanics/Latinos who took the Spanish version. A significant difference in the structure of the K6 for Hispanics/Latinos who took the English version as compared with the structure for Whites who took the English version would indicate that there is an effect of having a Hispanic/Latino ethnicity. However, this would not provide information about whether there is an effect of taking the Spanish version of the scale. A lack of significant difference in the structure of the K6 for Hispanics/Latinos who took the English version as compared with the structure for Whites who took the English version would indicate that there is not an effect of having a Hispanic/Latino ethnicity but that there is an effect of taking the Spanish version of the scale.

Constrained Baseline Analyses—Our initial constrained baseline IRT analyses indicated that all of the K6 items varied significantly across groups ($p < .05$) except for Items 3 (“restless”; $\chi^2[10] = 16.40$, $p = .09$) and 6 (“worthless”; $\chi^2[10] = 16.44$, $p = .09$). Given that these items were approximately equal in their consistency across groups, we decided to use Item 6 (“worthless”) as the referent in our free baseline IRT analyses to make our results more comparable with the separate analyses of race/ethnicity and language reported above.

Free Baseline Analyses—Table 7 reports the AIC, $-2 \log$ likelihood, and G^2 statistics for the free baseline IRT DIF analyses for the race/ethnicity by language combination. The baseline model only constrains Item 6 (“worthless”) to be consistent across groups. Each of the other models constrains Item 6 (“worthless”) as well as one other item to be consistent across groups. The tests reported in the last three columns in Table 7 determine whether constraining the second item leads to significantly worse fit relative to the baseline model. Results from DIF analyses show that Items 2 and 4 (“hopeless” and “depressed,” respectively) vary significantly across groups, but Items 1, 3, and 5 (“nervous,” “restless,” and “everything an effort,” respectively) do not. Our free baseline analyses do not provide us with a test of Item 6 (“worthless”), but we can conclude that it does not vary between groups based on the constrained baseline analyses.

In order to better understand the differences across the three racial/ethnic $\times$ language groups, effect sizes were examined to determine which groups had structures that were substantially different from each other. Results indicated that the structure found among Whites who took the English version was substantially different from both Hispanics/Latinos who took the English version (STDS = 2.697, ETSSD = 1.121) and Hispanics/Latinos who took the

Spanish version (STDS = 3.142, ETSSD = 1.104). Hispanics/Latinos who took the Spanish version were not notably different from Hispanics/ Latinos who took the English version (STDS = .232, ETSSD = .064). These results suggest that the previously observed effect of Spanish language may be an artifact resulting from its collinearity with race/ethnicity, but that the effect of race/ethnicity appears to be genuine.

Discussion

In our attempt to examine whether the K6 items function similarly or differently by race/ethnicity and/or language among diverse U.S. adults, we found clear evidence for measurement nonequivalence of the K6 by both race/ethnicity and language. The functioning of four items (“nervous,” “restless,” “depressed,” and “everything an effort”) varied by race/ethnicity and the functioning of four items (“nervous,” “hopeless,” “restless,” and “depressed”) varied by language. Half of the K6 items (i.e., 3 items, “nervous,” “restless,” and “depressed”) showed DIF by both race/ethnicity and language. The magnitude of DIF was large for the comparisons of non-Hispanic Whites versus Asians and Hispanics/Latinos, English versus Spanish, and the White English group versus the Hispanic/Latino English group and the Hispanic/Latino Spanish group, suggesting meaningful differences. This systematic DIF in the K6 scale suggests that the scale may have overestimated psychological distress for the Asian, Hispanic/Latino, and Spanish-speaking groups. This evident measurement nonequivalence in racially/ethnically and linguistically diverse populations in the United States raises concerns about the use of the K6 as a screening assessment for cross-cultural comparisons in national or international surveys.

The most striking finding was the general lack of measurement equivalence in the K6 across diverse racial/ethnic and linguistic groups. All K6 items (except for the referent item “worthless”) exhibited DIF by either race/ethnicity or language. Given the frequent use of the K6 in diverse populations, researchers should pay careful attention to cross-cultural comparability when interpreting results from the K6, especially when the K6 measure is used to provide prevalence or incidence rates of SPD among racially/ethnically or linguistically diverse adults. Our results showing different latent scores for optimal discrimination of psychological distress for Hispanics/Latinos, Asians, and the Spanish version in comparison with Whites and the English version suggest that using the standard cutoff score of 13 or higher on the K6 scale to determine SPD among people from diverse cultural and linguistic backgrounds may lead to misclassification. This differential misclassification could potentially result in overestimation of SPD for Hispanics/Latinos and Asians and Spanish-speaking individuals compared with non-Hispanic Whites and English-speaking individuals, respectively. Therefore, using the cutoff score of 13 or higher on the K6 may identify certain individuals as having SPD, who apparently do not.

Apparent measurement nonequivalence identified among Hispanics/Latinos and those interviewed in Spanish should be highlighted given the rapid growth of these population groups in the United States. All of the K6 items (except for the referent “worthless” item) showed evidence of DIF when assessed among Hispanics/Latinos, and all K6 items except “worthless” and “everything is an effort” showed evidence of DIF when assessed in Spanish.

DIF observed in the overall group for both the effects of race/ ethnicity and language was mostly contributed by DIF found among Hispanics/Latinos and those interviewed in Spanish. Although it is not clear why DIF in the K6 is more prominent among Hispanics/Latinos or individuals interviewed in Spanish than other racial/ethnic or linguistic groups, previous measurement equivalence research showed greater DIF among Hispanics/Latinos than among other racial/ethnic groups (e.g., Kim, Chiriboga, & Jang, 2009). In a measurement equivalence study of the Center for Epidemiologic Studies Depression Scale (CES-D) comparing older Whites, Blacks, and Mexican Americans, Kim et al. (2009) found that 80% of the depressive symptom items included in the CES-D scale (i.e., 16 items) showed DIF among Mexican Americans in comparison with Whites, and 45% of the CES-D items (i.e., 9 items) showed DIF among Mexican Americans in comparison with Blacks, whereas only 2 items showed DIF among Blacks in comparison with Whites. Kim et al. (2009) noted that depressive symptoms may be experienced, interpreted, and expressed differently among Mexican Americans in comparison with Whites and Blacks. Similar to the reasons suggested by Kim et al. (2009) and Kim (2010), culture-specific perception, interpretation, and reporting that are unique to the Hispanic/Latino culture might have contributed to the apparent measurement nonequivalence in symptoms of psychological distress. In addition, given that previous research suggests Hispanics/Latinos' greater tendency to somaticize their psychological distress (Fabrega, 1990; Norris, Arnau, Bramson, & Meagher, 2004), greater somatic symptom reporting among Hispanics/Latinos might have contributed to the current findings.

One novel contribution to the field is our attempt to separate the effects of race/ethnicity from language. The identified effect of race/ethnicity persisted when the interview language was matched, whereas the effect of Spanish language did not exist when the participants' race/ethnicity was matched. These findings are clear evidence of the more dominant effects of race/ethnicity over language on measurement nonequivalence in the K6. One plausible explanation would be that race/ethnicity may be related to fundamental differences in the ways people think and express their feelings, which was suggested by previous literature (Kim et al., 2009). Although the present study did not investigate the differential effects of race/ethnicity and language in other racial/ethnic minority groups except for Hispanics/Latinos due to the sample size issue, future research should further explore the relation.

DIF identified among Asians should be emphasized, as not many studies examined measurement equivalence among Asian Americans in comparison with other racial/ ethnic groups in the United States. Half of the K6 items ("nervous," "restless," and "everything is an effort") exhibited DIF among Asians in comparison with Whites. It is notable that most of the differing items are somatic symptom-related psychological distress items. Previous research reports Asians' greater tendency to express somatic symptoms than Whites (e.g., Farooq, Gahir, Okyere, Sheikh, & Oyebode, 1995). Also, there is a line of research suggesting that differences in somatic symptom severity may influence differences in symptom perception and reporting among Asians (e.g., Mak & Zane, 2004). Given the identified DIF for these somatic symptoms assessed by the K6, careful interpretation of scores for psychological distress and/or prevalence of SPD is needed when the K6 is used to identify SPD among Asians. Future research focusing on analyses of general somatic

symptoms could explain plausible reasons for the measurement nonequivalence of the K6 among Asians.

In addition to the aforementioned DIF, measurement equivalence observed in the present study deserves discussion. Overall, the “worthless” item was the only item that functioned equivalently across diverse racial/ethnic and linguistic groups in our sample. We observed that the functioning of the K6 items among African Americans/Blacks and AI/ANs was equivalent to that of Whites, and the functioning of the items in the Korean, Vietnamese, Mandarin, and Cantonese versions of the K6 scale were equivalent to the functioning of those items in English. This implies that researchers can make meaningful cross-cultural comparisons using the K6 scale among these racial/ethnic or linguistic groups. This also suggests that these culturally and linguistically diverse groups may perceive, interpret, and express their symptoms of psychological distress similarly. Given the wide use of the K6 nationally and internationally, the cross-cultural comparability of the K6 identified in the current investigation is encouraging. Validation research should be conducted by using nationally representative data and by expanding to other racial/ethnic or linguistic groups or to other countries.

Our findings have important implications for research and health policy. As mentioned earlier in the introduction, the K6 has been used mostly in large national surveys such as the NHIS, NHSDA, and Behavioral Risk Factor Surveillance System to assess psychological distress in the general population. Given that results from these national surveys are often used in federal reports, research using the K6 included in survey data may influence important mental health policy decisions. Thus, when researchers rely on the K6 to report psychological distress or SPD in diverse racial/ethnic or linguistic groups, they should be aware of the potential risk for misclassification among Hispanics/Latinos, Asians, or those completing the measure in Spanish, and should be careful before making conclusions from results using the K6.

Some potential limitations of the present study should be considered when interpreting the results. First and foremost, the generalizability of this study is limited by our use of data solely collected from California. Future investigations should examine the equivalence of the K6 across diverse samples using nationally representative data. Second, uneven distribution of racial/ethnic and linguistic groups may potentially limit the current study. Future investigation using more equally distributed diverse samples would be desirable. In-depth exploration for more than the linguistic and racial/ethnic minority groups assessed in the present study would enhance the quality of results. Third, other important factors such as age, gender, socioeconomic status (i.e., educational attainment and income), and acculturation that could potentially contribute to observed differences were not examined in the present study due to our primary focus on culture, race/ethnicity, and language. It may be possible that these demographic characteristics as well as interactions of these factors with race/ethnicity and/or language might affect differential responses to items to measure psychological distress. For example, given that a recent study found evidence of an age bias in the “fatigue” item included in the K6 (Sunderland, Hobbs, Anderson, & Andrews, 2012), future research should further examine the effects of other confounding variables on the measurement equivalence of the K6 scale. As another example, given the high rate of

intermarriage among Asians, especially in the West including California (Wang, 2012), different rates of intermarriage across diverse racial/ethnic groups might have contributed to response patterns observed in the present study. Fourth, racial/ethnic terms used in the survey may not represent their ethnocultural characteristics properly and/or do not seem parallel across different racial/ethnic groups in terms of capturing the geographic nature of the labels, which suggests that the field as a whole has a poor track record on proper labels. Fifth, our attempt to distinguish language from race/ethnicity could raise questions for contemporary approaches to culture in psychological science, and also opens questions for which language should be used in the development of alternative forms of the K6 or other screening instruments (e.g., Fassaert et al., 2009) as well as whether or not the Hispanics/Latinos who took the Spanish versus English versions of the K6 share key elements of ethnocultural background. Last, subgroup differences within each racial/ethnic or linguistic group were not investigated in the present study. Hispanics/Latinos, Asians, and Spanish-speaking individuals have heterogeneous characteristics with regard to nativity, country of origin, and number of years since immigration, which may have a significant impact on measurement equivalence of the K6 and should be explored in future research. For example, Crockett, Randall, Shen, Russell, and Driscoll (2005) found evidence that nonequivalence of a depressive symptom measure is influenced by cultural predictors varying by subgroups of Hispanic adolescents. In addition, given that the California sample includes recent Hispanic/Latino and Asian immigrants, researchers should consider testing the measurement equivalence of the K6 in international Hispanic/Latino and Asian samples compared with Hispanic/Latino and Asian immigrants in the United States to examine the effects of immigration-related characteristics. Further analyses on these subgroup differences in measurement equivalence may provide significant clinical implications.

Notwithstanding the limitations, the present study contributes to the literature by reporting a lack of cultural and linguistic equivalence in the K6 items among U.S. adults and by identifying the more dominant effects of race/ethnicity over language. Researchers using the K6 scale to screen for SPD should be aware that people from diverse racial/ethnic and linguistic backgrounds may not be screened equivalently with the K6 scale due to DIF. Using a biased measure for cross-cultural comparison has detrimental effects resulting in misdiagnoses, misclassification, and eventually mistreatment. Therefore, when the K6 scale is used to screen people from diverse racial/ethnic or linguistic groups (especially Hispanics/Latinos, Asians, or those speaking Spanish) for psychological distress, researchers should understand that diverse cultural and linguistic groups may experience and express symptoms of psychological distress differently. After testing for DIF in the K6 items with other populations from different national surveys, it is recommended that items that show consistently high magnitude DIF be removed, revised, or replaced. In addition, researchers may need to consider adjusting cutoff scores for certain racial/ethnic or linguistic groups, especially Asians, Hispanics/Latinos, and Spanish-speaking people.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study is supported by a grant (K01AG045342) funded by the National Institute on Aging.

References

- Andersen LS, Grimsrud A, Myer L, Williams DR, Stein DJ, Seedat S. The psychometric properties of the K10 and K6 scales in screening for mood and anxiety disorders in the South African Stress and Health study. *International Journal of Methods in Psychiatric Research*. 2011; 20:215–223. [PubMed: 22113964]
- Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*. 1981; 46:443–459.
- Bolt DM. A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*. 2002; 15:113–141.
- Cai, L., Thissen, D., du Toit, SHC. IRTPRO for Windows (Version 2.1) [Statistical software]. Lincolnwood, IL: Scientific Software; 2011.
- California Health Interview Survey. Sample design. n.d. Retrieved from <http://ucla-dev-web01.reliam.com/chis/design/Pages/sample.aspx>
- Crockett LJ, Randall BA, Shen YL, Russell ST, Driscoll AK. Measurement equivalence of the center for epidemiological studies depression scale for Latino and Anglo adolescents: A national study. *Journal of Consulting and Clinical Psychology*. 2005; 73:47–58. [PubMed: 15709831]
- Fabrega H. Hispanic mental health research: A case for cultural psychiatry. *Hispanic Journal of Behavioral Sciences*. 1990; 12:339–365.
- Farooq S, Gahir MS, Okyere E, Sheikh AJ, Oyeboode F. Somatization: A transcultural study. *Journal of Psychosomatic Research*. 1995; 39:883–888. [PubMed: 8636920]
- Fassaert T, De Wit MAS, Tuinebreijer WC, Wouters H, Verhoeff AP, Beekman ATF, Dekker J. Psychometric properties of an interviewer-administered version of the Kessler Psychological Distress scale (K10) among Dutch, Moroccan and Turkish respondents. *International Journal of Methods in Psychiatric Research*. 2009; 18:159–168. [PubMed: 19701920]
- Furukawa TA, Kawakami N, Saitoh M, Ono Y, Nakane Y, Nakamura Y, Kikkawa T. The performance of the Japanese version of the K6 and K10 in the World Health Survey Japan. *International Journal of Methods in Psychiatry Research*. 2008; 17:152–158.
- Furukawa TA, Kessler RC, Slade T, Andrews G. The performance of the K6 and K10 screening scales for psychological distress in the Australian National Survey of Mental Health and Well-Being. *Psychological Medicine*. 2003; 33:357–362. [PubMed: 12622315]
- Harremoës, P., Tusnády, G. Information divergence is more χ^2 -distributed than the χ^2 -statistics. *Information Theory Proceedings (ISIT)*, 2012 IEEE International Symposium; Cambridge, MA: IEEE; 2012. p. 533–537.
- Humes, K., Jones, NA., Ramirez, RR. Overview of race and Hispanic origin, 2010. U.S. Department of Commerce, Economics and Statistics Administration, U.S. Census Bureau; 2011. Retrieved from <http://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf>
- Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SLT, Zaslavsky AM, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine*. 2002; 32:959–976. [PubMed: 12214795]
- Kessler RC, Barker PR, Colpe LJ, Epstein JF, Gfroerer JC, Hiripi E, Zaslavsky AM. Screening for serious mental illness in the general population. *Archives of General Psychiatry*. 2003; 60:184–189. [PubMed: 12578436]
- Kim G. Measuring depression in a multicultural society: Conceptual issues and research recommendations. *Hallym International Journal of Aging*. 2010; 12:27–46.

- Kim G, Bryant AN, Parmelee P. Racial/ethnic differences in serious psychological distress among older adults in California. *International Journal of Geriatric Psychiatry*. 2012; 27:1070–1077. [PubMed: 22161758]
- Kim G, Chiriboga DA, Jang Y. Cultural equivalence in depressive symptoms in older White, Black, and Mexican American adults. *Journal of the American Geriatrics Society*. 2009; 57:790–796. [PubMed: 19484834]
- Kim G, Chiriboga DA, Jang Y, Lee S, Huang CH, Parmelee P. Health status of older Asian Americans in California. *Journal of the American Geriatrics Society*. 2010; 58:2003–2008. [PubMed: 20929469]
- Lee S, Tsang A, Ng KL, Ma YL, Guo W, Mak A, Kwok K. Performance of the 6-item Kessler scale for measuring serious mental illness in Hong Kong. *Comprehensive Psychiatry*. 2012; 53:584–592. [PubMed: 22104556]
- Mak WWS, Zane NWS. The phenomenon of somatization among community Chinese Americans. *Social Psychiatry and Psychiatric Epidemiology*. 2004; 39:967–974. [PubMed: 15583904]
- McVeigh KH, Galea S, Thorpe LE, Malsby C, Henning K, Sederer LI. The epidemiology of nonspecific psychological distress in New York City, 2002 and 2003. *Journal of Urban Health*. 2006; 83(3):394–405. [PubMed: 16739043]
- Meade AW. A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*. 2010; 95:728–743. [PubMed: 20604592]
- Mitchell CM, Beals J. The utility of the Kessler Screening Scale for Psychological Distress (K6) in two American Indian communities. *Psychological Assessment*. 2011; 23:752–761. [PubMed: 21534694]
- Norris MP, Arnau RC, Bramson R, Meagher MW. The efficacy of somatic symptoms in assessing depression in older primary care patients. *Clinical Gerontologist*. 2004; 27(1-2):43–57.
- Ponce NA, Lavarreda SA, Yen W, Brown ER, DiSogra C, Satter DE. The California Health Interview Survey 2001: Translation of a major survey for California's multiethnic population. *Public Health Reports*. 2004; 119:388–395. [PubMed: 15219795]
- Pratt LA. Serious psychological distress, as measured by the K6, and mortality. *Annals of Epidemiology*. 2009; 19:202–209. [PubMed: 19217003]
- Sakurai K, Nishi A, Kondo K, Yanagida K, Kawakami N. Screening performance of K6/K10 and other screening instruments for mood and anxiety disorders in Japan. *Psychiatry and Clinical Neurosciences*. 2011; 65:434–441. [PubMed: 21851452]
- Samejima, F. Estimation of latent ability using a response pattern of graded scores [Monograph supplement]. *Psychometrika*. 1969. Retrieved from <https://www.psychometricsociety.org/sites/default/files/pdf/MN17.pdf>
- Samejima, F. Graded response model. In: van der Linden, W., Hambleton, RK., editors. *Handbook of modern item response theory*. New York, NY: Springer; 1997. p. 85–100.
- Stark S, Chernyshenko OS, Drasgow F. Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*. 2006; 91:1292–1306. [PubMed: 17100485]
- Sunderland M, Hobbs MJ, Anderson TM, Andrews G. Psychological distress across the lifespan: Examining age-related item bias in the Kessler 6 Psychological Distress Scale. *International Psychogeriatrics*. 2012; 24:231–242. [PubMed: 21933463]
- Teresi JA, Ramirez M, Lai JS, Silver S. Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health. *Psychology Science Quarterly*. 2008; 50:538. [PubMed: 20165561]
- Wang, W. *The rise of intermarriage: Rates, characteristics vary by race and gender*. Washington, DC: Pew Research Center; 2012.
- Zagefka H. The concept of ethnicity in social psychological research: Definitional issues. *International Journal of Intercultural Relations*. 2009; 33:228–241.
- Zumbo, BD. *A handbook on the theory and methods of Differential Item Functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*.

Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense; 1999.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Characteristics of the CHIS Samples ($n = 44,846$).

	$M \pm SD$ or %						F (chi-square)
	Total ($n = 44,846$)	White ($n = 31,769$)	African American/ Black ($n = 1,950$)	Hispanic/Latino ($n = 5,753$)	Asian ($n = 4,874$)	AI/AN ($n = 500$)	
Language							(37070.74)***
English ($n = 39,918$)	89.0	99.8	99.8	47.6	62.0	99.2	
Spanish ($n = 3,075$)	6.9	0.2	0.2	52.4	0	0.8	
Vietnamese ($n = 885$)	2.0	0	0	0	18.1	0	
Korean ($n = 616$)	1.4	0	0	0	12.6	0	
Mandarin ($n = 187$)	0.4	0	0	0	3.8	0	
Cantonese ($n = 165$)	0.4	0	0	0	3.4	0	
Age (years)	56.12 $\pm$ 17.22	59.24 $\pm$ 16.31	54.52 $\pm$ 17.06	44.27 $\pm$ 16.28	50.56 $\pm$ 16.72	54.34 $\pm$ 15.94	1188.01***
Female	59.1	59.2	63.9	59.4	56.2	61.2	(36.89)***
Married	52.6	52.0	33.8	51.5	65.9	45.4	(639.85)***
Educational attainment							(7950.50)***
<High school	9.4	3.8	8.4	39.1	10.9	16.4	
High school diploma	21.8	20.6	24.5	28.2	19.7	29.0	
Some college	68.8	75.5	67.2	32.6	69.4	54.6	
Annual income (US\$)	72363.49 $\pm$ 64063.77	79749.59 $\pm$ 65968.19	53948.66 $\pm$ 51276.55	40070.52 $\pm$ 41008.41	71948.03 $\pm$ 65636.51	50496.47 $\pm$ 52557.78	551.75***
Total K6 (range: 0-24)	2.88 $\pm$ 3.50	2.73 $\pm$ 3.33	3.01 $\pm$ 3.89	3.54 $\pm$ 3.92	2.92 $\pm$ 3.60	4.01 $\pm$ 4.59	80.47***
Serious psychological distress (K6-13)	5.9	5.5	7.5	7.4	5.1	11.9	(81.37)***

Note. CHIS = California Health Interview Survey; AI/AN = American Indian/Alaska Native.

 $p < .001$.

Table 2

Free Baseline DIF Tests for Race/Ethnicity.

Model	-2 log likelihood	AIC	G ²	df	G ² difference from baseline	df Difference from baseline	G ² difference	p value
Baseline	399,235	399,511	11363.25	15,486	—	—	—	—
Item 1 (nervous)	399,454	399,690	11398.35	15,506	35.1	20	.02	.02
Item 2 (hopeless)	399,380	399,617	11387.38	15,506	24.13	20	.24	.24
Item 3 (restless)	399,677	399,912	11417.17	15,506	53.92	20	.0001	.0001
Item 4 (depressed)	399,495	399,731	11432.82	15,506	69.57	20	<.0001	<.0001
Item 5 (everything an effort)	399,529	399,765	11426.58	15,506	63.33	20	<.0001	<.0001

Note. DIF = differential item functioning; AIC = Akaike information criterion; df = degrees of freedom. Item 6 (worthless) was used as a reference item.

Table 3

STDS and ETSSD for Comparisons of the K6 Structure for Whites to the Structures Found in Other Racial/Ethnic Groups.

Effect size	Race/ethnicity (referent: White)			
	African American/Black	Hispanic/Latino	Asian	AI/AN
STDS	0.220	2.938	2.842	0.333
ETSSD	0.069	1.105	1.143	0.087

Note. STDS = Signed Test Difference in the Sample; ETSSD = Expected Test Score Standardized Difference; AI/AN = American Indian/Alaska Native. STDS represents the expected difference between the two groups in terms of the sum score for the scale. The ETSSD provides a similar measure rescaled so that it is on the same metric as Cohen's *d*.

Table 4

Free Baseline DIF Tests for Language.

Model	-2 log likelihood	AIC	G ²	df	G ² difference from baseline	df Difference from baseline	Chi-square <i>p</i>
Baseline	426,500	426,830	14348.73	15,459	—	—	—
Item 1 (nervous)	426,915	427,195	14386.47	15,484	37.74	25	.05
Item 2 (hopeless)	426,843	427,123	14384.83	15,484	36.10	25	.07
Item 3 (restless)	427,110	427,390	14388.65	15,484	39.92	25	.03
Item 4 (depressed)	426,895	427,175	14407.46	15,484	58.73	25	.0002
Item 5 (everything an effort)	426,985	427,265	14377.80	15,484	29.07	25	.26

Note. DIF = differential item functioning; AIC = Akaike information criterion; *df* = degrees of freedom. Item 6 (worthless) was used as a reference item.

Table 5

STDS and ETSSD for Comparisons of the K6 Structure for the English Version With the Structures Found in Other Language Versions.

Effect size	Language (referent: English)				
	Spanish	Vietnamese	Korean	Cantonese	Mandarin
STDS	3.165	-1.058	-0.637	-0.998	-1.083
ETSSD	1.093	-0.342	-0.170	-0.245	-0.325

Note. STDS = Signed Test Difference in the Sample; ETSSD = Expected Test Score Standardized Difference. STDS represents the expected difference between the two groups in terms of the sum score for the scale. The ETSSD provides a similar measure rescaled so that it is on the same metric as Cohen's *d*.

Table 6

Cross-Tabulation of the Relation Between Hispanic/ Latino Ethnicity and Spanish Language.

Language	Ethnicity	
	White (<i>n</i> = 31,767)	Hispanic/Latino (<i>n</i> = 5,753)
English (<i>n</i> = 34,454)	31,715 (84.53%)	2,739 (7.30%)
Spanish (<i>n</i> = 3,066)	52 (0.14%)	3,014 (8.03%)

Table 7

Free Baseline DIF Tests Comparing White English Version, Hispanic/Latino English Version, and Hispanic/Latino Spanish Version Groups.

Model	-2 log likelihood	AIC	G ²	df	G ² difference from baseline	df Difference from baseline	Chi-square <i>p</i>
Baseline	329,943	330,111	11323.98	15,540	—	—	—
Item 1 (nervous)	330,108	330,256	11335.86	15,550	11.88	10	.29
Item 2 (hopeless)	330,092	330,249	11350.89	15,550	26.91	10	.003
Item 3 (restless)	330,164	330,312	11336.96	15,550	12.98	10	.22
Item 4 (depressed)	330,221	330,369	11382.14	15,550	58.16	10	<.0001
Item 5 (everything an effort)	330,090	330,238	11338.22	15,550	14.24	10	.16

Note. DIF = differential item functioning; AIC = Akaike information criterion; *df* = degrees of freedom. Item 6 (worthless) was used as a reference item.