Published in final edited form as:

J Am Geriatr Soc. 2009 May ; 57(5): 790–796.

Cultural Equivalence in Depressive Symptoms in Older White, Black, and Mexican-American Adults

Giyeon Kim, PhD, David A. Chiriboga, PhD, and Yuri Jang, PhD

Department of Aging and Mental Health Disparities, Louis de la Parte Florida Mental Health Institute, University of South Florida, Tampa, Florida. Effective August 16, 2009, Giyeon Kim will be in the Department of Psychology and the Center for Mental Health and Aging at the University of Alabama.

Abstract

OBJECTIVES—To examine cultural equivalence in responses to depressive symptom items of three racial or ethnic elderly groups.

DESIGN—Cross-sectional analyses of two national data sets.

SETTING—The New Haven Established Populations for Epidemiologic Studies of the Elderly (EPESE) and the five-state Hispanic EPESE (H-EPESE).

PARTICIPANTS—Whites (n = 1,876) and blacks (n = 464) were drawn from the New Haven EPESE and Mexican Americans (n = 2,623) were drawn from the H-EPESE.

MEASUREMENT—The original 20-item version of the Center for Epidemiologic Studies Depression Scale (CES-D).

RESULTS—From differential item functioning analyses, a lack of measurement equivalence was found for 16 depressive symptom items. Mexican Americans were predisposed to endorse 12 depressive symptoms. Blacks were more likely than whites to endorse two interpersonal items (unfriendly and disliked). Mexican Americans were more likely than whites to respond to four positive affect items (good, hopeful, happy, and enjoyed) and more likely than blacks to endorse three positive items (good, happy, and enjoyed).

CONCLUSION—Results suggested response bias to depressive symptom items in racially and ethnically diverse older adults. Mexican Americans were more likely than whites to endorse the large number of depressive symptom items. Blacks were much more likely to respond in patterns similar to those of the whites. Findings from this study provide a foundation for developing culturally appropriate depression measures in health disparities research.

Keywords

cultural equivalence; depressive symptoms; CESD; measurement equivalence; differential item functioning

A growing body of literature documents racial and ethnic disparities in depressive symptoms.^{1,2} A number of cross-cultural and cross-national studies on depressive symptoms have found evidence that prevalence rates of probable depression vary dramatically across diverse racial/ and ethnic groups, with rates ranging from 1.5% to 32.0%.^{2–5} Because of such cultural variations, it has become a virtual truism in cross-cultural research that racially

and ethnically diverse groups manifest different prevalence rates of probable depression and different group means on standard inventories. One unresolved issue in comparing depressive symptoms across diverse population has been the equivalence of measures.^{6,7} Measurement equivalence is of particular importance, because if depressive symptom measures have differential meanings or validity across diverse cultural groups, group comparisons may be misleading and the estimated prevalence of depression inaccurate.⁶ It is often unclear whether actual differences in depression or differential item functioning (DIF) cause different depressive symptom scores across cultural groups. If people from different cultures but with equal levels of depression are more or less likely to endorse specific symptom items as a consequence of cultural grouping, those latter items will function differently across groups, they will show DIF.⁸ Measures containing DIF items may be invalid for between-group comparisons because their scores are indicative of attributes other than those that the test is intended to measure.⁹ Therefore, a high priority in assessing the cross-racial and -ethnic comparability of many widely used depression instruments should be to distinguish DIF (differential responses in the face of equivalent true scores on the latent trait) from impact (differences in item performance across groups due to real differences in the underlying target attribute). Focus in the present study was on the widely used Center for Epidemiologic Studies Depression Scale (CES-D).¹⁰ Since its development on samples of European Americans, the CES-D has been used in many cross-cultural studies. Despite its apparent usefulness in such applications,¹¹ there is evidence it has unique measurement properties across groups,^{5,6,12–15} although most of the studies investigating cross-cultural measurement differences have focused on subscale analyses. Given that cultural differences across diverse racial and ethnic groups may be related to unique patterns of response at the item level,¹⁶ these subscale analyses may not be enough to understand and determine unique psychometric properties of the CES-D.⁸ The potential of the CES-D items to function differentially across multiracial and -ethnic elderly groups thus becomes a priority in cross-cultural research.

A few studies have used DIF methods to investigate CES-D item bias according to race or ethnicity and found evidence of differential function.^{17–19} One DIF analysis conducted in white, Japanese, Native American, and Argentinean undergraduates found that whites were predisposed to endorse positive CES-D items and that Japanese and Argentineans were more likely to inhibit endorsement of the same positive items.¹⁸ Generalizations from this study were limited because of the use of a small and nonrepresentative sample. Testing DIF in representative samples of younger adults suggested that two of four positive feeling items (I felt hopeful about the future and I enjoyed life) showed that African Americans overendorsed DIF, whereas Hispanics tended to inhibit the expression of positive affect.¹⁹ Only two studies^{17,20} have investigated racial and ethnic item differences on the CES-D in older adults. Using the same New Haven Established Populations for Epidemiologic Studies of the Elderly (EPSE) data set employed in the present investigation, both studies found evidence that blacks were more likely than whites to endorse two interpersonal relation items (people are unfriendly and people dislike me). No study has fully considered and tested DIF items in the CES-D across three or more racial or ethnic elderly groups. Moreover, none have included older Mexican Americans, the largest subgroup of Hispanics in the United States.

The purpose of this study was to examine the cultural equivalence of the CES-D items across three racial and ethnic elderly groups: whites, blacks, and Mexican Americans. Specifically, the study focused on identifying race-and ethnicity-related DIF items in the CES-D that function differentially, as well as a core set of CES-D items that function equivalently across racially and ethnically diverse elderly groups.

METHODS

Sample

The New Haven EPESE provided the white and black samples and the Hispanic EPESE (H-EPESE) provided the Mexican-American sample. The New Haven EPESE is a longitudinal study of community-dwelling participants aged 65 and older collected in one of four geographic locations (East Boston, MA; New Haven, CT; Iowa and Washington counties, IA, and north-central North Carolina) and included whites and blacks at baseline (1982). The H-EPESE is a longitudinal study of Mexican Americans aged 65 and older from Texas, New Mexico, Colorado, Arizona, and California and was modeled after the design of the EPESE studies to compare with other populations in 1993/94. Using the first waves of the New Haven EPESE and H-EPESE, subjects were included in the analyses if they responded to all CES-D items (1,876 whites, 464 blacks, and 2,623 Mexican Americans).

Measures

The original 20-item version of the CES-D¹⁰ contains four positively stated items (reverse-coded) and 16 negatively stated items. The items ask how often symptoms were experienced during the previous week. Responses were rated on a 4-point scale, with categories presented in the following order: rarely or none of the time (0), some or a little of the time (1), much of the time (2), and most or all of the time (3). Total possible scores for 20 items ranged from 0 to 60, with higher scores indicating more-severe depressive symptoms. Scores of 16 or higher are typically viewed as evidence of probable depression.²¹ Reliability was satisfactory in the present sample: a α 0.86 for whites, a α 0.84 for blacks, and a α 0.88 for Mexican Americans.

Analytic Strategy

Researchers have generally used one of two methods that are capable of detecting DIF: multiple-group confirmatory factor analysis (CFA) and item response theory (IRT). Because these approaches use different algorithms and often generate different results, the present DIF analyses used both. Using both approaches has been advocated as a more-conservative approach to identifying DIF, with only those items identified by both being treated as DIF.²² Analyses began with a confirmatory analysis through LISREL 8.8 software (Scientific Software International, Inc., Lincoln-wood, IL) for structural equation modeling for the underlying unidimensionality of depressive symptoms. This was necessary, because the DIF detection methods used in this investigation assume that a single dominant factor underlies item responses.²³ Results from one-factor CFA supported unidimensionality, where goodness-of-fit indices for three racial or ethnic groups all exceeded 0.90, indicating generally adequate fits of the one-factor model to the data (for whites, comparative fit index (CFI) 0.95, normed fit index (NFI) 0.94, nonnormed fit index (NNFI) 0.95; for blacks, CFI 0.93, NFI 0.90, NNFI 0.93; and for Mexican Americans, CFI 0.91, NFI 0.90, NNFI 0.90).

After verifying unidimensionality, the application of IRT and CFA DIF detection using the likelihood ratio tests proceeded. In applying and interpreting the two methods, guidelines previously suggested²³ were followed, because that approach has demonstrated high power and low type I error rates across a wide variety of simulation conditions. In essence, it was suggested testing for DIF using a common strategy that can be implemented in CFA and IRT: a fully free-baseline model with strict Bonferroni corrected P-values for flagging DIF items.

CFA DIF Detection

CFA DIF analyses involving item loadings and intercepts were conducted using an analogous strategy with LISREL 8.8. Using the free-baseline model, in which only the parameters of the referent (Item 1) are constrained across groups, baseline and constrained models were run in succession, and the chi-square (w^2) difference statistics for the nested model comparisons were evaluated using a Bonferroni-corrected critical P-value. When the observed w^2 difference was greater than the corresponding critical value (Bonferroni corrected, $w^2 \geq 11.88$ with 2 degrees of freedom (df)), the item was flagged DIF.

IRT DIF Detection

Because the CES-D scale is polytomous, Samejima's Graded Response model was chosen, using the MULTILOG program (Scientific Software International, Inc.). For this model, each 4-category item has one discrimination parameter (a) and three location parameters (b_1 , b_2 , and b_3). The discrimination parameter reflects the extent to which an item differentiates between levels of underlying depression, and items with higher a are generally preferred because they are more informative in a psychometric sense. The location parameters refer to the point on the underlying depression scale at which the probability is 50% for endorsing the first category relative to the last three categories (b_1 : 0 vs 1, 2, 3), the first two categories relative to the last two categories (b_2 : 0, 1 vs 2, 3), and the first three categories relative to the fourth category (b_3 : 0, 1, 2 vs 3).

To assess model-data fit, the MODFIT program (Stark, <http://io.psych.uiuc.edu/irt/downloads.asp>) was used to fit the graded response model to the data. To determine good model-data fit, this study used adjusted w^2 to df ratios for item singles, doubles, and triples. Adjusted w^2 to df ratios for item singles, doubles, and triples all showed less than three, indicating good model-data fit.²⁴

The concurrent calibration method was subsequently used to put the reference and focal group parameters on a common metric with Item 1 as an anchor. In this step, whites (in two cases of white-black and white-Mexican American comparisons) and Mexican Americans (only in the case of Mexican American-black comparisons) were designated as the reference group, whose latent mean was set to 0. Mexican Americans (only in the case of white-Mexican American comparisons) and blacks (in two cases of white-black and Mexican American-black comparisons) were designated as the focal group; its latent mean was free to vary. As described for the CFA DIF method, the free-baseline model strategy was also used for each CES-D item, and differences in relative goodness of fit were examined with respect to critical w^2 statistics. Each w^2 difference was compared with Bonferroni-corrected P-values (corrected, $w^2 \geq 16.31$ with 4 df), and items exhibiting DIF were flagged.

RESULTS

Descriptive Information of Sample

As shown in Table 1, the white sample included more individuals aged 75 and older: 46.7% for whites and 32.3% for blacks and Mexican Americans. More than half were female for all three groups (56.6% of whites, 63.4% of blacks, and 58.1% of Mexican Americans). More Mexican Americans had less than an eighth-grade education (77%) than whites, the majority of whom had more than an eighth-grade education. Blacks were less likely to be married and had greater functional limitations than whites and Mexican Americans. The study variable, the CES-D, showed significantly different mean scores across the three racial and ethnic groups, with higher scores for Mexican Americans than for whites or blacks. Moreover, Mexican Americans (23.1%) consistently exhibited a greater likelihood for probable depression than levels reported for whites (16.0%) or blacks (14.4%) ($w^2 \geq 44.17$, $P < .001$).

Descriptive Item Statistics

As shown in Table 2, the 20 CES-D items were compared across the three racial and ethnic groups using the analysis of variance test. Significant mean differences were identified for 12 items (Items 3, 4, 6, 7, 9, 10, 12, 14, 15, 16, 17, 18, and 19). In each case, Mexican Americans consistently had higher mean scores than the other two groups, with an exception of two items (Items 15 and 19) on which blacks had higher means. These mean differences currently include effect (differences in item performance across groups due to real differences in depressive symptoms) and bias (differential responses in the face of equivalent true scores on depressive symptoms).

DIF Analyses

Table 3 summarizes results of CFA and IRT DIF analyses. In both analyses, item bias was tested at the equivalent total scores of depression across groups. As mentioned earlier, this study followed suggestions of other researchers,^{22,25,26} in focusing on DIF items detected using both methods. This was done as a rigorous strategy that would minimize possible type I error rates. In the CFA and IRT methods, Item 1 was used as a referent. Nineteen model comparisons were made, and each of the CES-D items was constrained to be equal across groups in each model.

As shown in Table 3, three group comparisons were made (whites vs blacks, whites vs Mexican Americans, and Mexican Americans vs blacks). White–Mexican American group comparisons exhibited the greatest number of DIF (16 items) and white–black group comparisons flagged the fewest number of DIF (2 items). Across all three group comparisons, only one item (Item 15, people are unfriendly) consistently exhibited DIF, and 16 of 20 items exhibited DIF at least once. Of these 16 DIF items, Mexican Americans had a greater propensity to endorse 11 items (Items 2, 3, 5, 6, 7, 9, 10, 13, 17, 18, and 20) than whites or blacks. Blacks were more likely to endorse two interpersonal problem items (Items 15, people are unfriendly, and 19, people dislike me) than whites and Mexican Americans. Only four items (Items 1, bothered by things; 2, poor appetite; 11, restless sleep; and 14, lonely) showed no evidence of DIF, suggesting that they functioned equivalently across three racial or ethnic groups.

Whites Versus Blacks

In the comparison of whites and blacks, CFA identified five DIF items (Items 12, 15, 16, 17, and 19), and IRT identified four DIF items (Items 7, 8, 15, and 19). Thus, two DIF items (Items 15, people are unfriendly, and people dislike me) were identified in both methods. The same findings have been previously observed in two DIF studies,^{17,20} and the present study supported their findings using the same samples drawn from the New Haven EPESE but different DIF methods (the Mantel-Haenszel¹⁷ and the multiple indicators, multiple causes model²⁰). As was the case in these two studies, blacks were more likely than whites to endorse the two interpersonal items.

Whites Versus Mexican Americans

In the white–Mexican American comparison, CFA identified 16 DIF items, and IRT identified 19 DIF items. IRT also identified all 16 of the CFA-flagged items (Items 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 15, 16, 17, 18, 19, and 20). In other words, 80% of the CES-D items functioned differently in whites and Mexican Americans. All four positive affect items (Items 4, feel good as others; 8, feel hopeful; 12, happy; and 16, enjoy life) exhibited DIF in both approaches, with Mexican Americans being more likely to endorse these items. Mexican Americans were more likely to endorse 12 DIF items assessing negative affect.

Mexican Americans Versus Blacks

In the comparison of Mexican Americans and blacks, CFA identified nine DIF items, and IRT identified 17 items. There were nine common DIF items (Items 3, 4, 5, 6, 12, 15, 16, 17, and 18), suggesting that nearly half of the CES-D items functioned differently for Mexican Americans and blacks. Three of the four positive affect items (Items 4, feel good as others; 12, happy; and 16, enjoy life) showed DIF, with Mexican Americans being more likely to endorse these items. This finding parallels those for the comparisons with whites and suggests that Mexican Americans may be more likely to report positive feelings than blacks or whites. Mexican Americans were also more likely to endorse all negative affect items exhibiting DIF.

DISCUSSION

The primary goal of this study was to investigate the cultural equivalence of the CES-D in older whites, blacks, and Mexican Americans. This study may be the first to examine item bias in the CES-D across racially and ethnically diverse elderly populations that include older Mexican Americans. DIF analyses indicated that, in whites, blacks, and Mexican Americans, 80% of the CES-D items functioned differently on at least one comparison. That left only four items of the 20 CES-D items as being identified to function similarly in all three racial and ethnic groups. The bottom line is that all three groups clearly did not report their symptoms of depression equivalently, a finding that emphasizes the need for further study of measurement equivalence in at least this depression-screening instrument.

The most striking finding in the present DIF analyses was the general lack of measurement equivalence of the CES-D in the comparison of Mexican Americans with whites and blacks. Of the 16 items with DIF for the white–Mexican American comparison, Mexican Americans were predisposed to endorse all of the 16 symptoms (including four positive affect items). The comparison with blacks was nearly as dramatic, with nearly half of the items on the CES-D manifesting item bias and the results indicating higher levels of endorsement by Mexican Americans.

In the comparison of whites and blacks, results replicated previous findings from two published DIF studies on the CES-D items in older adults that used the same data set employed in the present analysis.^{17,20} Blacks consistently overendorsed the only two interpersonal relation items on the CES-D (people are unfriendly and people dislike me), which may reflect perceptions of racial discrimination by blacks. It has been well documented that blacks generally experience more-disadvantaged social conditions than whites and are more generally likely to report racial discrimination.^{27,28} Disproportionate responses to the two interpersonal items may reflect a confounding of depressive symptoms with perceived racial prejudice, a possibility that has been raised in other published studies with blacks,³ although these studies were not testing DIF.

The greater tendency to endorse depressive symptoms in Mexican Americans was of particular interest, because DIF analyses with this population are rare. Previous research indicating that Mexican Americans are less hesitant than non-Hispanic whites to admit their symptoms of psychological distress may partially explain the findings.^{16,29} This response style may help to explain their relatively high depression scores. Overall, these results suggest that using the standard cutoff scores of 16 or higher with Mexican American elderly people may lead to misclassification, resulting in high false-positive rates. Future work is warranted on the appropriate cutoff scores for the CES-D in Mexican-American population.

One intriguing finding was that Mexican Americans were much more likely to endorse positive affect items than the other two groups, suggesting that Hispanics tend to exaggerate

at least the reporting of positive feelings such as feeling happy and enjoyed life. Being less hesitant to express feelings may also explain this item bias in the CES-D among Mexican Americans. This finding did not parallel results reported in two studies of young adults,^{18,19} that reported that young Hispanics were less likely than whites to endorse positive affect items. Further research should be done regarding cohort differences on positive feeling items among Hispanics. Consistent with previous studies,^{17,20} whites and blacks showed a similar response tendency in expressing their positive feelings, indicating that they may share values, attitudes, and beliefs regarding the expression of positive feelings.

Another finding of interest in the present investigation was that Mexican Americans appeared to report somatic symptoms differently than did whites and blacks. Mexican Americans were more likely to endorse four of the seven items described in the literature as presenting somatic symptoms in the CES-D (Items 5, trouble concentrating; 7, everything is an effort; 13, talk less; and 20, cannot get going). They were also more likely to endorse one somatic item (Item 5, trouble concentrating) in the Mexican American–black comparison. The literature has suggested that Hispanics are more likely to somatize their psychological distress,^{30,31} and the current findings partially support this, although the current study identified three of the four items without DIF (bothered by things, poor appetite, and restless sleep) as somatic symptom items. Given that this study involved only older adults and that older adults in general may tend to somatize their depressive symptoms,³² an age effect may have reduced racial or ethnic differences in somatic symptom items in this sample.

This was the first study using two popular DIF methods, CFA and IRT, to detect DIF with the CES-D. When this joint strategy was applied to detect DIF, comparisons of whites and blacks showed perfect agreement with two previous DIF studies^{17,20} that used the same sample employed here. This suggests that the combined use of the two statistical approaches increases the accuracy for testing measurement equivalence of depressive symptom inventories; using multiple DIF detection approaches has been recommended for more-accurate results.^{22,25,26} However, given that the CFA and IRT showed some discrepancy in the identification of DIF items, the results raise questions as to how to interpret DIF items that only one method detects. Although the question remains unanswered, careful interpretation should be made of these uncommonly identified DIF items across different methods for future research, and further investigation will be needed to find a source of these discrepant items.

With respect to study limitations, one factor that was not controlled in the study was the potential influence of historical time and cohort differences between the samples from the New Haven EPESE and the H-EPESE. The New Haven EPESE was collected in 1981/82, whereas the H-EPESE was collected in 1993/94. The more than 10-year difference between those two samples may have led to differential response patterns. In addition, the present study included a relatively small sample of blacks and thus used the unequal sample sizes of the three groups for DIF analyses, which may need careful attention. Both limitations underscore the importance of appropriate nationally representative data sets that can provide enough information to capture racial and ethnic disparities in health.

Despite the abovementioned limitations, findings from this study hold important implications for research, practice, and public policies. When the CES-D is used to screen for depression, researchers and clinicians should be aware of the risk that elderly people from different cultural backgrounds may tend to be misclassified, leading directly to under- or overdiagnosis of depression. Use of inaccurate measures could also lead to misguided public policies. Therefore, in light of the consequences of using nonequivalent measures, researchers should pay careful attention to making measures more reliable and culturally

appropriate, as well as to establishing measurement equivalence of the existing depression measures, which is the first and crucial step before divergent groups are compared.

These findings also provide a caution with respect to the interpretation of scores on standard screening tools for depression. Clinicians should recognize that depressive symptoms may present differently across different cultural groups. For example, when clinicians assess depression in older Mexican Americans, they need to adjust their own concepts of depression to permit appropriate diagnosis and treatment for Mexican Americans by recognizing their greater tendency to express symptoms of depression. Clinicians need to become more culturally competent so that they can incorporate some cultural concepts such as self-orientation and family values when needed.

In conclusion, the present study highlights the importance of considering depressive symptoms that diverse cultural groups may experience and express differently. Current item analyses provided evidence of Mexican Americans' predisposition to endorse depressive symptoms more than other cultural groups with the same levels of depression, suggesting that adjustment such as modification of culturally inappropriate items and changes in cutoff scores may be warranted. More work remains to be done, especially with regard to understanding potential sources of response bias, such as sociodemographic characteristics, as well as identifying magnitude of DIF for practical significance. Future research should be conducted using diverse populations of elderly people with and without clinical diagnoses of depression to validate these findings. Ultimately, this avenue of research may lead to the development of a screening tool that is as free of item bias as possible across diverse racial and ethnic groups.

REFERENCES

1. Coyne JC, Marcus SC. Health disparities in care for depression possibly obscured by the clinical significance criterion. *Am J Psychiatr* 2006;163:1577–1579. [PubMed: 16946183]
2. Dunlop DD, Song J, Lyons JS, et al. Racial/ethnic differences in rates of depression among pre-retirement adults. *Am J Public Health* 2003;93:1945–1952. [PubMed: 14600071]
3. Blazer DG, Landerman LR, Hays JC, et al. Symptoms of depression among community-dwelling elderly African-American and White older adults. *Psychol Med* 1998;28:1311–1320. [PubMed: 9854272]
4. Gonzalez HM, Haan MN, Hinton L. Acculturation and the prevalence of depression in older Mexican Americans: Baseline results of the Sacramento Area Latino Study on Aging. *J Am Geriatr Soc* 2001;49:948–953. [PubMed: 11527487]
5. Foley KL, Reed PS, Mutran EJ, et al. Measurement adequacy of the CES-D among a sample of older African-Americans. *Psychiatr Res* 2002;109:61–69.
6. Crockett LJ, Randall BA, Shen YL, et al. Measurement equivalence of the Center for Epidemiological Studies Depression Scale for Latino and Anglo adolescents: A national study. *J Consult Clin Psychol* 2005;73:47–58. [PubMed: 15709831]
7. Liang, J. Assessing cross-cultural comparability in mental health among older adults.. In: Skinner, JH.; Teresi, JA.; Holmes, D., et al., editors. *Multicultural Measurement in Older Populations*. Springer Publishing Company; New York: 2002. p. 11-21.
8. Teresi, JA. Statistical methods for examination of differential item functioning (DIF) with applications to cross-cultural measurement of functional, physical and mental health.. In: Skinner, JH.; Teresi, JA.; Holmes, D., et al., editors. *Multicultural Measurement in Older Populations*. Springer Publishing Company; New York: 2002. p. 23-34.
9. Perkins AJ, Stump TE, Monahan PO, et al. Assessment of differential item functioning for demographic comparisons in the MOS SF-36 health survey. *Qual Life Res* 2006;15:331–348. [PubMed: 16547771]
10. Radloff L. The CES-D Scale: A self-report depression scale for research in the general population. *Appl Psychol Meas* 1977;1:385–401.

11. Mui, AC.; Burnette, D.; Chen, LM. Cross-cultural assessment of geriatric depression: A review of the CES-D and GDS.. In: Skinner, JH.; Teresi, JA.; Holmes, D., et al., editors. *Multicultural Measurement in Older Populations*. Springer Publishing Company; New York: 2002. p. 147-177.
12. Callahan CM, Wolinsky FD. The effect of gender and race on the measurement properties of the CES-D in older adults. *Med Care* 1994;32:341–356. [PubMed: 8139299]
13. Miller TQ, Markides KS, Black SA. The factor structure of the CES-D in two surveys of elderly Mexican Americans. *J Gerontol B Psychol Sci Soc Sci* 1997;52B:259–269.
14. Nguyen HT, Kitner-Triolo M, Evans MK, et al. Factorial invariance of the CESD in low socioeconomic status African Americans compared with a nationally representative sample. *Psychiatr Res* 2004;126:177–187.
15. Roberts RE, Vernon SW, Rhoades HM. Effects of language and ethnic status on reliability and validity of the Center for Epidemiologic Studies-Depression Scale with psychiatric patients. *J Nerv Ment Dis* 1989;177:581–592. [PubMed: 2677237]
16. McHorney CA, Fleishman JA. Assessing and understanding measurement equivalence in health outcome measures. *Med Care* 2006;44(Suppl 3):S205–S210. [PubMed: 17060829]
17. Cole SR, Kawachi I, Maller SJ, et al. Test of item-response bias in the CES-D scale: Experience from the New Haven EPESE study. *J Clin Epidemiol* 2000;53:285–289. [PubMed: 10760639]
18. Iwata N, Buka S. Race/ethnicity and depressive symptoms: A cross-cultural/ ethnic comparison among university students in East Asia, North and South America. *Soc Sci Med* 2002;55:2243–2252. [PubMed: 12409137]
19. Iwata N, Turner RJ, Lloyd DA. Race/ethnicity and depressive symptoms in community-dwelling young adults: A differential item functioning analysis. *Psychiatr Res* 2002;110:281–289.
20. Yang FM, Jones RN. Center for Epidemiologic Studies-Depression scale (CES-D) item response bias found with Mantel-Haenszel method was successfully replicated using latent variable modeling. *J Clin Epidemiol* 2007;60:1195–1200. [PubMed: 17938063]
21. Andresen EM, Malmgren JA, Carter WB, et al. Screening for depression in well older adults: Evaluation of a short form of the CES-D. *Am J Prev Med* 1994;10:77–84. [PubMed: 8037935]
22. Hambleton RK. Good practices for identifying differential item functioning. *Med Care* 2006;44(Suppl 3):S182–S188. [PubMed: 17060826]
23. Stark S, Chernyshenko OS, Drasgow F. Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *J Appl Psychol* 2006;91:1292–1306. [PubMed: 17100485]
24. Drasgow F, Levine MV, Tsien S, et al. Fitting polytomous item response theory models to multiple-choice tests. *Appl Psychol Meas* 1995;19:143–165.
25. Schaffer BS, Riordan CM. A review of cross-cultural methodologies for organizational research: A best-practice approach. *Organ Res Methods* 2003;6:169–215.
26. Wang M, Russell SS. Measurement equivalence of the Job Description Index across Chinese and American workers: Results from confirmatory factor analysis and item response theory. *Educ Psychol Meas* 2005;65:709–732.
27. Ren XS, Amick BC, Willimans DR. Racial/ethnic disparities in health: The interplay between discrimination and socioeconomic status. *Ethn Dis* 1999;9:151–165. [PubMed: 10421078]
28. Williams DR. The health of U.S. racial and ethnic population. *J Gerontol B Psychol Sci Soc Sci* 2005;60B(Special Issue II):53–62. [PubMed: 16251591]
29. Haberman P. Ethnic differences in psychiatric symptoms reported in community surveys. *Public Health Rep* 1970;85:495–502. [PubMed: 4987769]
30. Angel R, Guarnaccia PJ. Mind, body, and culture: Somatization among Hispanics. *Soc Sci Med* 1989;28:1229–1238. [PubMed: 2660279]
31. Fabrega H. Hispanic mental health research: A case for cultural psychiatry. *Hispanic J Behav Sci* 1990;12:339–365.
32. Norris MP, Arnau RC, Meagher MW, et al. The efficacy of somatic symptoms in assessing depression in older primary care patients. *Clin Gerontol* 2005;27:43–57.

Table 1

Descriptive Characteristics of the Sample

Characteristic	New Haven EPESE (N 5 2,340)		Hispanic EPESE (N 5 2,623)	
	Whites (n 5 1,876)	Blacks (n 5 464)	Mexican Americans (n 5 2,623)	F (chi-square)
Aged >75, %	45.7	32.3	32.3	(89.6***)
Female, %	56.6	63.4	58.1	(7.0*)
Educational attainment, %				(1,279.3***)
<eighth grade	23.6	50.5	77.0	
Eighth–11th grade	40.5	31.9	13.3	
12th grade	20.1	13.2	6.7	
>12th grade	15.8	4.4	3.1	
Married, %	45.9	29.4	56.1	(66.1***)
Activity of daily living limitations, %				(45.7***)
0	86.2	79.9	87.1	
>1	13.8	20.1	12.9	
CES-D score, mean $\pm$ standard deviation	8.0 $\pm$	7.9 $\pm$ 7.8	12.9 10.1 $\pm$ 9.3	33.8***
Probable depression (CES-D score >16), %	8.4 16.0	14.4	23.1	(44.2***)

Table 2

Center for Epidemiologic Studies Depression Scale (CES-D) Scores in Whites, Blacks, and Mexican Americans

CES-D Item	Mean $\pm$ Standard Deviation			
	Whites	Blacks	Mexican American	
(Possible Range of Scores in Each Item: 0–3)	(n 5 1,876)	(n 5 464)	(n 5 2,623)	F
1. I was bothered by things that usually don't bother me.	0.44 $\pm$ 0.81	0.36 $\pm$ 0.79	0.43 $\pm$ 0.78	1.6
2. I did not feel like eating; my appetite was poor.	0.35 $\pm$ 0.78	0.43 $\pm$ 0.85	0.35 $\pm$ 0.71	2.6 [†] ,
3. I felt that I could not shake off the blues even with help from my family or friends.	0.33 $\pm$ 0.73	0.28 $\pm$ 0.66	0.37 $\pm$ 0.74	4.4 [*] , [§] ,
4. I felt that I was just as good as other people.	0.29 $\pm$ 0.77	0.26 $\pm$ 0.71	0.87 $\pm$ 1.21	203.7 ^{***} , [§] ,
5. I had trouble keeping my mind on what I was doing.	0.40 $\pm$ 0.76	0.34 $\pm$ 0.67	0.43 $\pm$ 0.75	2.8 ^k
6. I felt depressed.	0.50 $\pm$ 0.81	0.44 $\pm$ 0.76	0.54 $\pm$ 0.81 ^{*,k}	3.6
7. I felt everything was an effort.	0.51 $\pm$ 0.90	0.55 $\pm$ 0.96	0.61 $\pm$ 0.92	7.6 ^{***} , [§]
8. I felt hopeful about the future. [†]	0.93 $\pm$ 1.22	0.99 $\pm$ 1.23	0.94 $\pm$ 1.16	0.5
9. I thought my life had been a failure.	0.20 $\pm$ 0.60	0.20 $\pm$ 0.61	0.26 $\pm$ 0.64	5.3 ^{**} , [§] ,
10. I felt fearful.	0.27 $\pm$ 0.63	0.29 $\pm$ 0.68	0.32 $\pm$ 0.66	4.3 [*] , [§]
11. My sleep was restless.	0.59 $\pm$ 0.98	0.50 $\pm$ 0.86	0.55 $\pm$ 0.90	1.9
12. I was happy.	0.62 $\pm$ 0.99	0.52 $\pm$ 0.89	0.85 $\pm$ 1.07	39.5 ^{***} , [§] ,
13. I talked less than usual.	0.35 $\pm$ 0.79	0.36 $\pm$ 0.79	0.39 $\pm$ 0.75	1.4
14. I felt lonely.	0.50 $\pm$ 0.87	0.48 $\pm$ 0.84	0.47 $\pm$ 0.83	0.7
15. People were unfriendly.	0.20 $\pm$ 0.59	0.37 $\pm$ 0.79	0.25 $\pm$ 0.68	13.5 ^{***} , [†] , [§] ,
16. I enjoyed life. [†]	0.46 $\pm$ 0.91	0.34 $\pm$ 0.75	0.91 $\pm$ 1.12	139.0 ^{***} , [†] , [§] ,
17. I had crying spells.	0.21 $\pm$ 0.58	0.14 $\pm$ 0.50	0.42 $\pm$ 0.78	66.5 ^{***} , [§] , ^k
18. I felt sad.	0.44 $\pm$ 0.72	0.43 $\pm$ 0.73	0.54 $\pm$ 0.83	10.5 ^{***} , [§] ,
19. I felt people disliked me.	0.11 $\pm$ 0.43	0.25 $\pm$ 0.64	0.18 $\pm$ 0.53	19.0 ^{***} , [†] , [§] ,
20. I could not get going.	0.34 $\pm$ 0.71	0.34 $\pm$ 0.69	0.35 $\pm$ 0.73	1.3

[†] Reverse-coded item.

[†] A significant mean difference between whites and blacks was obtained at the .05 level.

[§] A significant mean difference between whites and Hispanics was obtained at the .05 level.

^k A significant mean difference between blacks and Hispanics was obtained at the .05 level.

* P<.05

** P<.01

*** P<.001.

Table 3
Differential Item Functioning (DIF) Results from Confirmatory Factor Analysis (CFA) and Item Response Theory (IRT) Methods

Chi-Square (Difference)						
Whites Versus Mexican Americans Whites Versus Blacks Americans Versus Blacks						

Model	CFA * (Δ df 5 2)	IRT [†] (Δ df 5 4)	CFA * (Δ df 5 2)	IRT [†] (Δ df 5 4)	CFA * (Δ df 5 2)	IRT [†] (Δ df 5 4)
Baseline model (referent: Item 1)	2,033.9	27,259.1	6,344.6	58,833.0	5,495.8	44,396.8
Comparison models						
2. Appetite	10.8	15.6	0.9	80.3 ^{DIF}	9.87	33.3 ^{DIF}
3. Blues	8.4	13.4	18.2 ^{DIF}	18.4 ^{DIF}	19.51 ^{DIF}	25.8 ^{DIF}
4. Good [‡]	8.8	3.5	893.0 ^{DIF}	346.2 ^{DIF}	307.16 ^{DIF}	124.0 ^{DIF}
5. Mind	5.3	7.8	12.9 ^{DIF}	81.4 ^{DIF}	14.62 ^{DIF}	30.8 ^{DIF}
6. Depressed	6.8	7.2	17.6 ^{DIF}	39.7 ^{DIF}	13.57 ^{DIF}	29.6 ^{DIF}
7. Effort	5.7	17.2 ^{DIF}	52.9 ^{DIF}	60.1 ^{DIF}	4.52	52.8 ^{DIF}
8. Hopeful [‡]	10.8	17.8 ^{DIF}	41.3 ^{DIF}	212.2 ^{DIF}	1.68	72.4 ^{DIF}
9. Failure	4.9	5.0	32.6 ^{DIF}	33.2 ^{DIF}	7.8	6.7
10. Fearful	4.5	7.2	31.3 ^{DIF}	52.8 ^{DIF}	0.8	15.3
11. Sleep	9.6	10.1	6.0	80.8 ^{DIF}	2.0	37.3 ^{DIF}
12. Happy [‡]	14.8 ^{DIF}	5.8	214.4 ^{DIF}	225.4 ^{DIF}	177.4 ^{DIF}	53.9 ^{DIF}
13. Talked less	0.4	4.3	12.7 ^{DIF}	80.1 ^{DIF}	2.1	25.4 ^{DIF}
14. Lonely	1.2	6.2	7.0	29.9 ^{DIF}	1.8	18.2 ^{DIF}
15. Unfriendly	29.4 ^{DIF}	41.5 ^{DIF}	34.5 ^{DIF}	24.3 ^{DIF}	15.7 ^{DIF}	43.0 ^{DIF}
16. Enjoyed [‡]	33.7 ^{DIF}	10.7	619.8 ^{DIF}	341.2 ^{DIF}	267.8 ^{DIF}	97.3 ^{DIF}
17. Crying	21.9 ^{DIF}	5.0	383.2 ^{DIF}	105.7 ^{DIF}	294.0 ^{DIF}	49.1 ^{DIF}
18. Sad	0.2	11.6	77.9 ^{DIF}	50.6 ^{DIF}	20.9 ^{DIF}	38.2 ^{DIF}
19. Disliked	47.7 ^{DIF}	48.5 ^{DIF}	146.1 ^{DIF}	72.6 ^{DIF}	10.8	38.5 ^{DIF}
20. Get going	1.4	7.4	17.2 ^{DIF}	74.3 ^{DIF}	6.3	29.4 ^{DIF}
Total number of DIF items	5	4	16	19	9	17

Note: Items in bold are common DIF items across CFA and IRT methods in each cross-racial or -ethnic comparison.

* In CFA, DIF flagged if chi-square (χ^2) was >11.88.

† In IRT, DIF flagged if χ^2 was >16.31.

‡ Reverse-coded item.