Published in final edited form as:
J Nerv Ment Dis. 2008 April ; 196(4): 297–306. doi:10.1097/NMD.0b013e31816a490e.

Differential Item Functioning Between Ethnic Groups in the Epidemiological Assessment of Depression

Joshua Breslau, PhD, ScD[*], **Kristin N. Javaras, DPhil**[†], **Deborah Blacker, MD, ScD**[‡,§], **Jane M. Murphy, PhD**[‡,§], and **Sharon-Lise T. Normand, PhD**[†,‖]

[*]Center for Reducing Health Disparities, Department of Internal Medicine, University of California, Davis School of Medicine, California

[†]Department of Biostatistics, Harvard School of Public Health

[‡]Department of Epidemiology, Harvard School of Public Health

[§]Department of Psychiatry, Massachusetts General Hospital

[‖]Health Care Policy, Harvard Medical School, Boston, Massachusetts

Abstract

A potential explanation for the finding that disadvantaged minority status is associated with a lower lifetime risk for depression is that individuals from minority ethnic groups may be less likely to endorse survey questions about depression even when they have the same level of depression. We examine this possibility using a nonparametric item response theory approach to assess differential item functioning (DIF) in a national survey of psychiatric disorders, the National Comorbidity Survey. Of 20 questions used to assess depression symptoms, we found evidence of DIF in 3 questions when comparing non-Hispanic blacks with non-Hispanic whites and in 3 questions when comparing Hispanics with non-Hispanic whites. However, removal of the questions with DIF did not alter the relative prevalence of depression between ethnic groups. Ethnic differences do exist in response to questions concerning depression, but these differences do not account for the finding of relatively low prevalence of depression among minority groups.

Keywords

Depression; measurement; bias; ethnicity; epidemiology

Although risk of depression is consistently associated with low socioeconomic status (Lorant et al., 2003), epidemiological studies have not found elevated risk for depression among socially disadvantaged minority groups in the United States. In fact, all surveys that have assessed depression in national samples of American adults have found significantly lower lifetime risk for depression among non-Hispanic blacks compared with non-Hispanic whites (Blazer et al., 1994; Breslau et al., 2006; Somervell et al., 1989; Weissman et al., 1991; Williams et al., 2007). In addition, no surveys have found higher risk for depression among Hispanics relative to non-Hispanic whites, whereas some have found significantly lower risk among Hispanics than non-Hispanic whites (Blazer et al., 1994; Breslau et al., 2006; Karno et al., 1987). However, because of the difficulty of establishing the cultural equivalence of

Send reprint requests to Joshua Breslau, PhD, ScD, Center for Reducing, Health Disparities, Department of Internal Medicine, University of California, Davis School of Medicine, 2921 Stockton Blvd., Sacramento, CA, 95817. E-mail: jabreslau@ucdavis.edu.

psychological constructs (Butcher et al., 2003; Janca, 2005; Room et al., 1996), including depression (Kleinman, 2004), concerns about the validity of this counterintuitive finding remain (Rogler, 1999). Measurement bias resulting from differential misclassification, with minorities being less likely to meet DSM criteria despite similar levels of underlying disorder, remains a potential methodological explanation for these findings (Rogler et al., 2001).

In this article, we investigate whether measurement bias in the assessment of depression can account for the unexpectedly low lifetime prevalence of depression among Hispanics and non-Hispanic blacks relative to non-Hispanic whites. Data come from the National Comorbidity Survey (NCS), a nationally representative survey of psychiatric disorders in which respondents were interviewed face-to-face by nonclinician interviewers using the Composite International Diagnostic Interview (CIDI), a fully structured diagnostic interview schedule. In the CIDI, depression was assessed in 2 stages: in the screening stage, questions regarding low mood and anhedonia were asked of the entire sample, and in the diagnostic stage, 20 questions regarding the 7 additional criterion symptoms for depression were asked only of those respondents who endorsed at least 1 question at the screening stage.

Potential Sources of Ethnic Differences in Depression

Four pathways through which ethnic differences in survey responses to a CIDI depression question might arise are illustrated in Figure 1. In the figure, these pathways are illustrated using the question concerning feelings of worthlessness, which is used to assess the criterion symptom of self-reproach.

In scenario A there are real differences in the prevalence of depression between ethnic groups. This leads to lower prevalence of the self-reproach symptom and consequently lower endorsement of the question about worthlessness. According to scenario A, we would expect similar group differences in responses to all questions. In scenario B, responses to the worthlessness question differ between groups because depression manifests differently between ethnic groups. Under this scenario, the symptom of self-reproach has a different relationship with depression depending on group membership. This leads to group differences in response to the worthlessness question but also to similar differences in other questions used to assess the self-reproach symptom, such as the question assessing excessive guilt. In scenario C, responses to the worthlessness question differ by group because ethnic groups differ in a general factor that affects their responses to all survey questions. For instance, groups may differ in their willingness to disclose potentially embarrassing personal information. Under this scenario, there would be similar group differences in responses to most or all questions. Finally, in scenario D, responses to the worthlessness question differ by group because ethnic groups differ in the way they interpret this particular question. Under this scenario, there would be group differences in response to the worthlessness question, but no differences in response to other questions that were interpreted consistently between groups.

Differential Item Functioning (DIF)

In all 4 scenarios, there would be observed differences between groups in the prevalence of depression. However, only in scenario A would these observed differences correspond to differences between groups in true underlying levels of depression. In contrast, for scenarios B to D, these observed differences would correspond not to differences between groups in the true underlying levels of depression, but instead to differences between groups in the relationship between the true underlying level of depression and responses to particular depression questions. In the psychometric literature, this latter phenomenon is known as DIF (Camilli and Shepard, 1994; Holland and Wainer, 1993). Simply put, DIF refers to any scenario where the probability of endorsing a particular question about depression differs for individuals who have the same underlying level of depression but belong to different groups.

DIF is illustrated graphically in Figure 2 using the item characteristic curves (ICC) for the worthlessness item among non-Hispanic whites and non-Hispanic blacks (Thissen et al., 1993). The ICC is a function that relates the probability that a respondent says "yes" to an item to his or her underlying level of the trait being measured. In the figure, the horizontal axis indicates the respondent's underlying level of depression, an unobserved (i.e., latent) variable which is estimated from the entire group of questions. The vertical axis indicates the respondent's likelihood of answering yes when asked the question about feelings of worthlessness. The curves represent the likelihoods among non-Hispanic black and non-Hispanic whites of endorsing feelings of worthlessness at each level of depression. In Figure 2, DIF is indicated by the gap between the 2 curves. Moreover, the impact of DIF on the assessment of depression can be inferred from the relative locations of the ICCs. Because the ICC for blacks is shifted to the right relative to the ICC for whites in the figure, blacks are less likely to endorse this question than whites with the same level of depression. This particular type of DIF is referred to as DIF against blacks and may lead to the underestimation of depression for blacks relative to whites. If the ICC for blacks were shifted to the left of that for whites, this would indicate that blacks are more likely to endorse the question than whites with the same level of depression. This type of question is said to exhibit DIF in favor of blacks and may lead to overestimation of depression for blacks relative to whites.

DIF and Bias

The existence of DIF indicates that the measurement of depression is also sensitive to some secondary factor, as in scenarios B, C, and D (Shealy and Stout, 1993a,b). However, the existence of this secondary factor is not necessarily problematic. "Benign DIF" (Douglas et al., 1996) occurs when the secondary dimension is auxiliary to the underlying trait of depression. For instance, in scenario B DIF may reflect important cultural differences in the experience of depression (Alegria and McGuire, 2003). On the other hand, "adverse DIF" (Douglas et al., 1996) occurs when the secondary factor is not relevant to the underlying trait of depression, as in scenarios C and D. In these scenarios, group differences in response are not relevant to the assessment of group differences in depression and can lead to erroneous conclusions about them.

Psychometric methods for detecting DIF using item response theory (IRT) have been developed and extensively applied in studies of achievement tests (Hambleton and Swaminatha, 1985; Hambleton et al., 1991; Holland and Wainer, 1993). An IRT-based assessment of DIF between cultural groups in the Center for Epidemiological Studies-Depression scale (CES-D) found that positively phrased questions function differently between cultural groups (Iwata and Buka, 2002). However, IRT-based methods have not been used to investigate DIF between ethnic groups in data from a diagnostic survey instrument designed to assess DSM-defined depression in the general population. In this report, we use a nonparametric approach to testing for DIF in the CIDI depression questions.

METHODS

Sample

As described in detail elsewhere (Kessler et al., 1994), the National Comorbidity Survey (NCS), conducted in 1990–1992, is based on a stratified, multistage area probability sample of persons aged 15 to 54 years of age in the noninstitutionalized civilian population in the 48 coterminous states. The response rate was 74%. The data are then weighted to account for the sample design, to adjust for nonresponse (based on a supplemental nonresponse survey), and to poststratify the sample to the US national population (Little et al., 1997).

Ethnicity was determined using 2 questions. The first question asks whether or not the respondent is Hispanic, and the second asks the respondent's race. For our analysis, 3 groups were defined. Hispanics were those who indicated Hispanic ethnicity regardless of race (n = 784). Blacks were those non-Hispanics who indicated black race (n = 929). Whites were those non-Hispanics who indicated white race (n = 6099). All others were excluded from this analysis.

Assessment of Depression

Respondents were interviewed with the CIDI (Kessler et al., 1998). The instrument was designed to assess psychiatric disorders according to the DSM-III-R definitions. The CIDI was found to have good reliability and validity for assessing depression in the WHO CIDI Field Trials (Wittchen, 1994). The instrument and data are available for public use at:http://webapp.icpsr.umich.edu/cocoon/SAMHDA-DISPLAY/06693.xml.

The CIDI obtained information on depressive episodes in 2 stages. The first stage consisted of screening questions that assessed whether the respondent had ever experienced a period of depressed mood or loss of interest in activities (anhedonia) of sufficient duration to warrant suspicion of a history of major depressive episode. Respondents who screened positive for either of these 2 symptoms advanced to the second stage where the remaining 7 symptoms of depression were assessed through a series of 20 yes/no questions; respondents who screened negative were not asked further about depression. In the second stage, each of the 7 remaining symptoms was assessed by between 1 and 4 questions. The diagnostic algorithm used to diagnose depression from these questions counted a symptom as present if 1 or more questions were endorsed for that symptom. Endorsed questions that may have resulted from organic causes and those that did not cluster together within a two-week time period were excluded. Respondents were diagnosed with depression if they had 5 or more symptoms present after the organic and clustering exclusions.

In our analysis, we examined responses to all 20 of the questions from the second stage and to the anhedonia question from the first stage. Responses were corrected for organic causes and clustering in an episode.

Detecting DIF

We used a nonparametric approach to the assessment of DIF as implemented in the SIBTEST statistical software package (for a technical description of the SIBTEST procedure, see Sheely and Stout, 1992a,b and Liang and Stout, 1998). SIBTEST estimates respondents' depression levels by counting the number of questions with which respondents agree. These estimates, and the test for DIF based on them, are nonparametric in the sense that they make no assumptions about the form taken by the ICCs or the underlying distribution of depression levels. The former is an important advantage for the depression questions in the NCS because there is evidence that the ICCs for questions from psychological scales do not follow parametric IRT models such as the 2-and 3-parameter Rasch models (Meijer and Baneke, 2004).

We first used SIBTEST to look for DIF in the questions. Because we did not have a priori information that allowed us to identify a subset of DIF-free depression questions, we adopted the "leave-one-out" approach to estimate depression levels. More specifically, we used the leave-one-out symptom count, which is the number of symptoms, other than the one being measured by the question being examined, that are present. We used the leave-one-out symptom count, rather than the leave-one-out question count, to address the dependence between questions measuring the same symptom (Wainer, 1995). Because we could not assume that the leave-one-out symptom count was uncontaminated by DIF, we used SIBTEST to perform 1-sided tests (for DIF against the minority group) in an iterative procedure because,

as we explain in the discussion section below, doing so guards against confusing DIF with group differences in depression levels when pervasive DIF is present. The other reason for using 1-sided tests was that only DIF against a minority group can result in lower prevalence, and our goal was to determine whether DIF could explain minority groups' lower prevalence of depression.

Our procedure for using SIBTEST to detect DIF was as follows. For each question, we tested the null hypothesis of no DIF against the alternative hypothesis of DIF against the minority group. If there were fewer than 5 minority or 5 white respondents with a particular value of the leave-one-out symptom count, then we required SIBTEST to omit that value from its calculations. If the *p* value from the 1-sided test was significant at the 0.05 level according to Benjamini and Hochberg's (1995) rules for controlling the false discovery rate with multiple tests, the question was deemed to exhibit significant DIF against the minority group. After a first pass through all of the questions, we removed any question found to have significant DIF against the minority group, provided it was not the sole question measuring a symptom, and then retested the remaining questions for DIF. This procedure was repeated until no questions with DIF remained, except for questions that were the sole question measuring a symptom. Finally, for questions found to have significant DIF, we attempted to distinguish between benign DIF due to group differences in the manifestation of depression and adverse DIF due to measurement bias by considering how many of the questions measuring the same symptom had DIF.

We also used SIBTEST to look for DIF at the symptom level because DIF that is not detectable at the question level may accumulate between questions measuring the same symptom and thereby impact depression scores (Wainer and Kiely, 1987; Wainer et al., 1991). We performed 1-sided tests for DIF at the symptom level by treating each symptom as a binary question that equaled 1 ("present") if the respondent answered yes to at least 1 question measuring that symptom and 0 ("absent") if no questions measuring that symptom were endorsed. Here, too, the leave-one-out symptom count was used to estimate depression levels.

The above purification procedures were performed first with blacks as the minority group, and then again with Hispanics as the minority group.

Calculating Lifetime Prevalence of Depression

We used the diagnostic algorithm described above to calculate the lifetime prevalence of depressive episodes for whites, blacks, and Hispanics, first using all of the questions and then using only the questions remaining after removing questions identified as having DIF in the purification procedure described above. We used the chi-square test to compare groups' lifetime prevalence of depressive episodes, before and after removing questions with DIF. The chi-square tests were adjusted for the complex survey design using the Taylor series linearization method implemented in SUDAAN (RTI, 2002).

RESULTS

Screening

Fifty-seven percent of the sample (n = 4501) screened positive for suspected lifetime history of depressive episode. The proportion screening positive was significantly lower (χ^2_1=7.9, p = 0.008) among blacks (51%; n = 505) than among whites (58%; n = 3585), but there was no significant difference (χ^2_1=0.35, p = 0.558) between Hispanics (57%; n = 411) and whites.

Question Endorsement and Symptom Prevalence

The proportions of the sample and of each ethnic group endorsing each depression question and meeting criteria for each symptom are presented in Table 1. Note that because these proportions are marginal, between-group differences do not necessarily imply DIF, which refers to between-group differences in proportions conditional on the true underlying level of depression. There were significant differences in endorsement proportions between blacks and whites. Blacks endorsed all but 2 of the 21 questions less frequently than whites, significantly so for 9 of the questions. Blacks had lower prevalence than whites for all 8 symptoms, and these differences reached statistical significance for 5 symptoms. Differences between Hispanics and whites were less pervasive. Hispanics were less likely than whites to endorse about half (10) of the questions and these differences reached significance in 2 cases, both in questions assessing suicidality. Hispanics differed from whites in the prevalence of only 1 symptom, suicidality. SIBTEST Analysis

Black-White Comparison

The results from the SIBTEST analyses with blacks as the minority group are presented in Table 2. Bold font is used to indicate questions and symptoms whose SIBTEST *p* value is significant at the $p = 0.05$ level using the Benjamini and Hochberg correction for multiple testing. To illustrate the interpretation of these results, we consider the statistics for the initial test of the worthlessness question in Table 2. The positive estimate of 0.064 indicates that blacks are, on average, 6.4% less likely to answer yes to the worthlessness question than whites with the same level of depression, and the bolded *p* value indicates that this difference is statistically significant.

In the initial test (left hand columns of Table 2), DIF was found for 3 questions: “lack of energy,” “felt worthless,” and “thoughts of suicide.” The direction of DIF in all 3 cases was against blacks, consistent with bias toward underestimation of depression among blacks relative to whites. At the symptom level, DIF was found for 2 of the symptoms: “loss of energy” and “self-reproach.”

The felt worthless and thoughts of suicide questions were removed in the second set of tests for DIF. The lack of energy question was not removed because it was the sole question available to assess the DSM symptom of loss of energy and therefore could not be removed without reducing the validity of the instrument. No new questions with DIF were identified in the second pass (right hand columns of Table 2).

The lifetime prevalence of depressive episode among whites and blacks was 18.6% and 12.6%, respectively ($\chi^2_1=10.69$, $p = 0.002$), using all the questions and 17.7% and 12.4%, respectively ($\chi^2_1=8.20$, $p = 0.007$), after the adjustment for DIF.

Hispanic-White Comparison

The results from the SIBTEST analyses with Hispanics as the minority group are presented in Table 3. In the initial pass (left hand columns of Table 3), the questions “gained weight,” “waking early,” and thoughts of suicide were significant, but at the symptom level DIF was found only for suicidality. In the second pass all 3 questions with DIF were removed, and no further questions were found to have DIF (right hand columns of Table 3). Using all the questions the prevalence of depressive episodes among whites and Hispanics was 18.6% and 18.3%, respectively ($\chi^2_1=0.02$, $p = 0.895$). After adjustment for DIF the prevalence of depressive episodes in both groups was 18.3% ($\chi^2_1=0.01$, $p = 0.994$).

DISCUSSION

In this national survey, both Hispanics and non-Hispanic blacks tended to endorse depression questions less frequently than non-Hispanic whites. This was particularly true for blacks, who were significantly less likely than whites to have 5 of the 8 symptoms examined. Hispanics were less likely than whites to have 1 of the 8 symptoms, suicidality. We used a nonparametric IRT approach to examine whether the lower levels of endorsement among minorities in this survey could be explained by DIF against minorities, defined as a lower probability of endorsing questions among minorities compared with non-Hispanic whites after adjusting for differences in underlying levels of depression.

DIF against minorities was found for several of the questions used to assess depression. DIF against minorities at the question level led to DIF against minorities at the symptom level for 2 symptoms in the black-white comparison and for 1 symptom in the Hispanic-white comparison. However, the existence of DIF does not necessarily imply that the conclusions regarding the relative prevalence of depression between groups is incorrect. To assess the impact of the DIF against minorities on estimates of the relative prevalence of depression between ethnic groups, we followed a procedure of removing questions with DIF against minorities if this could be done without adversely affecting the construct validity of the test. After performing this procedure, we were able to reduce DIF at the symptom level for the black-white comparison and eliminate DIF at the symptom level for the Hispanic-white comparison. Recalculating the prevalence of depression without the questions with DIF, we found no evidence that differences in lifetime prevalence between groups were significantly affected by DIF in the original questions. With respect to the main question posed in this study, we found no evidence that the low prevalence of depression among minorities relative to whites is a result of DIF.

DIF Against Blacks

Two questions were removed because they had significant DIF against blacks: felt worthless and thoughts of suicide. In both cases, there were other questions measuring the same symptom that were not found to exhibit DIF. This suggests that nuisance factors related to the specific wording of the question with DIF, rather than factors related more generally to the symptom, account for the observed DIF in both cases, which implies that the DIF in these 2 cases is likely to be adverse. Because of DIF in the felt worthless question, DIF was also found for the symptom of self-reproach. DIF in this symptom was no longer found when the felt worthless question was removed. This suggests that one means of improving the consistency of the assessment of depression would be to remove the felt worthless question. However, an equally valid and perhaps more clinically defensible approach would be to add an additional question with an alternative phrasing of the question that might be more commonly endorsed by minority groups at the same level of depression. Using questions with countervailing DIF by design may be the best way to achieve consistent measurement given the reality of cultural variations in the idioms of depression and the way that the questions are combined to determine a diagnostic assessment.

Although 1 of the 4 questions used to assess suicidality displayed DIF against blacks, no DIF was found at the level of the suicidality symptom even with this question included. This is most likely a result of the facts that (a) this question has a lower prevalence in the population than other questions used to assess this symptom and (b) blacks were slightly (nonsignificantly) more likely than whites to endorse other suicide questions at the same level of depression. Therefore, DIF in this question has a negligible impact on the relative prevalence of depression.

DIF against blacks was also found for the lack of energy question. Because this was the only question used to assess this symptom, we could not remove it from the test and maintain the

test's construct validity. Therefore, these data do not provide sufficient evidence to suggest whether DIF in this case is due to the specific wording of the question or to some clinically meaningful variation in the manifestation of depression between ethnic groups.

DIF Against Hispanics

Three questions were found to have DIF against Hispanics. Two of these questions, weight gain and early waking, did not result in DIF at the symptom level because of countervailing influences from other questions. Both of these questions inquire about discrete experiences that are understood to be instances of the more general symptom: weight gain is a discrete type of appetite and weight change, and early waking is a discrete type of sleep disturbance. Therefore, in neither case can we rule out the possibility that DIF for these questions is benign, due to an auxiliary factor that is clinically meaningful. Therefore, we cannot make a recommendation as to whether these questions should be removed or changed on the basis of these data.

The third question with DIF against Hispanics was the thoughts of suicide question. As in the black-white comparison, DIF in this question against Hispanics did not adversely affect consistency of measurement at the symptom level.

Limitations

These results should be interpreted in light of 4 limitations of the IRT approach to assessment of DIF and response bias. First, our assessment of DIF relies on an internal criterion to assess respondents' underlying levels of depression. Because we did not have a valid subset of questions, we had to use the leave-one-out approach to measure depression levels. This approach may result in estimates of depression that are contaminated by DIF, which creates problems with detecting DIF in individual questions and symptoms. In particular, it is impossible to detect DIF if all questions (or symptoms) have DIF of similar magnitude and direction (Camilli, 1993). This theoretical concern has been borne out in simulation studies by Gierl et al. (2004), who investigated the performance of SIBTEST in the presence of pervasive DIF. They found that in these circumstances, SIBTEST may fail to detect questions with significant DIF and, further, may falsely flag questions that in reality lack DIF.

Two aspects of our DIF detection procedure limited the possibility that our conclusions would be distorted by pervasive DIF. First, we used 1-sided tests to examine DIF against minorities. This prevented us from erroneously removing any DIF-free questions that, because of contamination in our underlying measure of depression, appeared to have DIF in favor of minorities. Second, we used an iterative purification procedure whereby we eliminated questions with DIF against minorities and then re-examined the remaining questions for DIF until no new questions with DIF were found (Camilli and Shepard, 1994). Because questions with severe DIF against minorities will still seem to have significant DIF and will therefore be removed after the first pass, the estimate of depression used in the second pass will be less contaminated by DIF. This means that we will then be able to detect questions with moderate DIF against minorities in the second pass, and so on and so forth. If most or all questions have similar amounts of DIF against minorities (e.g., because of a greater reluctance to disclose potentially embarrassing information), then we will not be able to detect DIF even with an iterative purification procedure based on 1-sided tests. However, in less extreme situations, the 1-sided iterative procedure will allow us to locate those questions that truly have DIF against the minority group.

A second limitation of this analysis is SIBTEST's assumption that DIF is unidirectional. This means that when DIF against minorities occurs in a question, minorities have lower probability of endorsing that question at all levels of depression. However, there maybe questions where

DIF occurs in opposite directions at different levels of depression. If this is the case, then our SIBTEST procedure would not allow us to detect these questions as having DIF.

A third limitation stems from the 2-stage structure of the CIDI assessment of depression, which does not collect full information on depressive symptoms for respondents who are negative on the initial screening questions. This means that we cannot investigate DIF in the screening questions where it may actually do the most damage in terms of biasing group prevalence estimates. For example, we still do not know whether blacks were less likely to screen positive for depression because of DIF in the screening questions or because of actual differences in prevalence. It would be valuable to repeat this analysis in a sample that contained full data on the screen negatives.

This study also has the limitation of sample size. The sample size limits our ability to examine DIF with respect to specific ethnic subgroups, such as Puerto Rican versus Mexicans in comparison with non-Hispanic whites. It also limits our ability to examine DIF within sociodemographic subgroups, such as males and females. Although our main conclusions support the findings of epidemiological comparisons between broad ethnic groups (e.g., Hispanics compared with non-Hispanic whites), it is important to recognize that neither this epidemiological pattern nor our methodological result necessarily holds true for all subgroups within the broad ethnic groups examined.

Finally, this study relies on retrospective recall of depressive symptoms. Though this is a potential limitation of the method, there is no evidence of differential recall between ethnic groups (Shrout et al., 1993).

CONCLUSIONS

Results of this analysis support 2 apparently conflicting conclusions. On the one hand, we found that some aspects of depression assessment differ significantly between ethnic groups. Specifically, differences were found with respect to questions assessing the symptoms of self-reproach, suicidality, lack of energy, weight gain, and sleep disturbance. Removing questions with DIF may not be the best way to address these measurement problems in future instruments. Another potential strategy is to include additional questions with countervailing differences in functioning between groups, provided that the measurement characteristics of questions within particular groups can be reliably determined. Although the fully structured research interviews used in epidemiological surveys differ from clinical interviews, these findings may also have clinical implications. Clinicians should be aware that assessments of particular aspects of depression maybe affected by choice of the idiom used in a clinical interview (Kleinman, 2004; Neighbors et al., 1999; Schmaling and Hernandez, 2005).

On the other hand, correction for these differences in the way people responded to survey questions did not change the epidemiological conclusions. When we removed the questions that contributed to underestimation of depression among minorities and recalculated the prevalence estimates, the original findings were not altered. This result increases our confidence in the consistent finding that socially disadvantaged ethnic minority status is not associated with higher risk for depression in the United States, a finding that has been consistent in epidemiological studies since the early 1980s. The analysis, however, cannot rule out measurement differences that are consistent across the entire range of depression questions or that affect the screening questions about anhedonia and depressed mood. Further, it cannot rule out nonmeasurement sources of bias such as survey nonresponse.

Acknowledgments

This research was partially supported by grants K01 MH66057-02 (to J.B.) and R01-MH54693 (to S.-L.T.N.) from the National Institute of Mental Health and by the NIH Training Program in Psychiatric Epidemiology and Biostatistics, grant number 5 T32 MN17119-22 (to K.N.J.).

REFERENCES

Alegria M, McGuire T. Rethinking a universal framework in the psychiatric symptom-disorder relationship. J Health Soc Behav 2003;44:257–274. [PubMed: 14582307]

Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc 1995;57:289–300.

Blazer DG, Kessler RC, McGonagle KA, Swartz MS. The prevalence and distribution of major depression in the general population: Results from the National Comorbidity Survey. Am J Psychiatry 1994;151:979–986. [PubMed: 8010383]

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

Butcher JN, Cheung FM, Lim J. Use of the MMPI-2 with Asian populations. Psychol Assess 2003;15:248–256. [PubMed: 14593825]

Camilli, G. The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues?. In: Holland, PW.; Wainer Hillsdale, H., editors. Differential Item Functioning. New Jersey: Lawrence Erlbaum Associates Publishers; 1993. p. 397-418.

Camilli, G.; Shepard, LA. Methods for Identifying Biased Test Items. Thousand Oaks: Sage Publications; 1994.

Douglas J, Roussos L, Stout W. Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their DIF. J Educ Meas 1996;33:465–484.

Gierl MJ, Gotzmann A, Boughton KA. Performance of SIBTEST when the percentage of DIF items is large. Applied Measurement in Education 2004;17(3):241–264.

Hambleton, RK.; Swaminatha, H. Item Response Theory: Principles and Applications. Boston: Kluwer-Nijhoff Publishers; 1985.

Hambleton, RK.; Swaminathan, H.; Jane Rogers, H. Fundamentals of Item Response Theory. Newbury Park: Sage Publications; 1991.

Holland, PW.; Wainer, H., editors. Differential Item Functioning. Hillsdale (NJ): Lawrence Erlbaum Associates, Publishers; 1993.

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]

Janca A. Diagnosis, classification and prevalance of somatoform disorders in a cross-cultural perspective. Aust N Z J Psychiatry 2005;39:A109–A109.

Karno M, Hough RL, Burnam MA, Escobar JI, Timbers DM, Santana F, Boyd JH. Lifetime prevalence of specific psychiatric disorders among Mexican Americans and non-Hispanic whites in Los Angeles. Arch Gen Psychiatry 1987;44:695–701. [PubMed: 3498453]

Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S, Wittchen HU, Kendler KS. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey. Arch Gen Psychiatry 1994;51:8–19. [PubMed: 8279933]

Kessler RC, Wittchen H-U, Abelson JM, Mcgonagle K, Schwarz N, Kendler KS, Knäuper B, Zhao S. Methodological studies of the Composite International Diagnostic Interview (CIDI) in the United States. Int J Methods Psychiatr Res 1998;7:33–55.

Kleinman A. Culture and depression. N Engl J Med 2004;351:951–953. [PubMed: 15342799]

Liang H, Stout W. Improved type I error control and reduced estimation bias for DIF detecting using SIBTEST. J Educ Behav Stat 1998;23:291–322.

Little RJ, Lewitsky S, Heeringa S, Lepkowski J, Kessler RC. Assessment of weighting methodology for the National Comorbidity Survey. Am J Epidemiol 1997;146:439–449. [PubMed: 9290504]

Lorant V, Deliege D, Eaton W, Robert A, Philippot P, Ansseau M. Socioeconomic inequalites in depression: A meta-analysis. Am J Epidemiol 2003;157:98–112. [PubMed: 12522017]

Meijer RR, Baneke JJ. Analyzing psychopathology items: A case for nonparametric items response theory modeling. Psycholo Methods 2004;9:354–368.

Neighbors HW, Trierweiler SJ, Munday C, Thompson EE, Jackson JS, Binion VJ, Gomez J. Psychiatric diagnosis of African Americans: Diagnostic divergence in clinician-structured and semistructured interviewing conditions. J Natl Med Assoc 1999;91:601–612. [PubMed: 10641496]

Rogler LH. Methodological sources of cultural insensitivity in mental health research. Am Psychol 1999;54:424–433. [PubMed: 10392472]

Rogler LH, Mroczek DK, Fellows M, Loftus ST. The neglect of response bias in mental health research. J Nerv Ment Dis 2001;189:182–187. [PubMed: 11277355]

Room R, Janca A, Bennett LA, Schmidt L, Sartorius N. WHO cross-cultural applicability research on diagnosis and assessment of substance use disorders: An overview of methods and selected results. Addiction 1996;91:199–220. [PubMed: 8835277]

RTI. Software for Survey Data Analysis (SUDAAN), Version 8.1. Research Triangle Park (NC): Research Triangle Institute; 2002.

Schmaling KB, Hernandez DV. Detection of depression among low-income Mexican Americans in primary care. J Health Care Poor Underserved 2005;16:780–790. [PubMed: 16311498]

Shealy, R.; Stout, W. An item response theory model for test bias and differential test functioning. In: Holland, PW.; Wainer, H., editors. Differential Item Functioning. Hillsdale (NJ): Lawrence Earlbaum; 1993a. p. 197-240.

Shealy R, Stout W. A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. Psychometrica 1993b;58:159–194.

Shrout PE, Canino GJ, Bird HR, Rubio-Stipec M, Bravo M, Burnam MA. Mental-health status among Puerto-Ricans, Mexican-Americans and non-Hispanic whites—the case of the Misbegotten Hypothesis—Response. Am J Community Psychol 1993;21:389–395.

Somervell PD, Leaf PJ, Weissman MM, Blazer DG, Bruce ML. The prevalence of major depression in black and white adults in five United States communities. Am J Epidemiol 1989;130:725–735. [PubMed: 2788995]

Thissen, D.; Steinberg, L.; Wainer, H. Detection of differential item functioning using the parameters of item response models. In: Holland, PW.; Wainer, H., editors. Differential Item Functioning. Hillsdale (NJ): Lawrence Erlbaum Associates, Publishers; 1993. p. 67-113.

Wainer H. Precision and differential item functioning on a testlet-based test: The 1991 law school admissions test as an example. Appl Meas Educ 1995;8:157–186.

Wainer H, Kiely GL. Item clusters and computerized adaptive testing: A case for testlets. J Educ Meas 1987;24:185–201.

Wainer H, Sireci SG, Thissen D. Differential testlet functioning: Definition and detection. J Educ Meas 1991;28:197–219.

Weissman, MM.; Bruce, ML.; Leaf, PJ.; Florio, LP.; Holzer, C. Affective disorders. In: Robins, LN.; Regier, DA., editors. Psychiatric Disorders in America: The Epidemiologic Catchment Area Study. New York: The Free Press; 1991.

Williams DR, Gonzalez HM, Neighbors H, Nesse R, Abelson JM, Sweetman J, Jackson JS. Prevalence and distribution of major depressive disorder in African Americans, Caribbean blacks and non-Hispanic whites: Results from the National Survey of American life. Arch Gen Psychiatry 2007;64:305–315. [PubMed: 17339519]

Wittchen HU. Reliability and validity studies of the WHO—Composite International Diagnostic Interview (CIDI): A critical review. J Psychiatr Res 1994;28:57–84. [PubMed: 8064641]

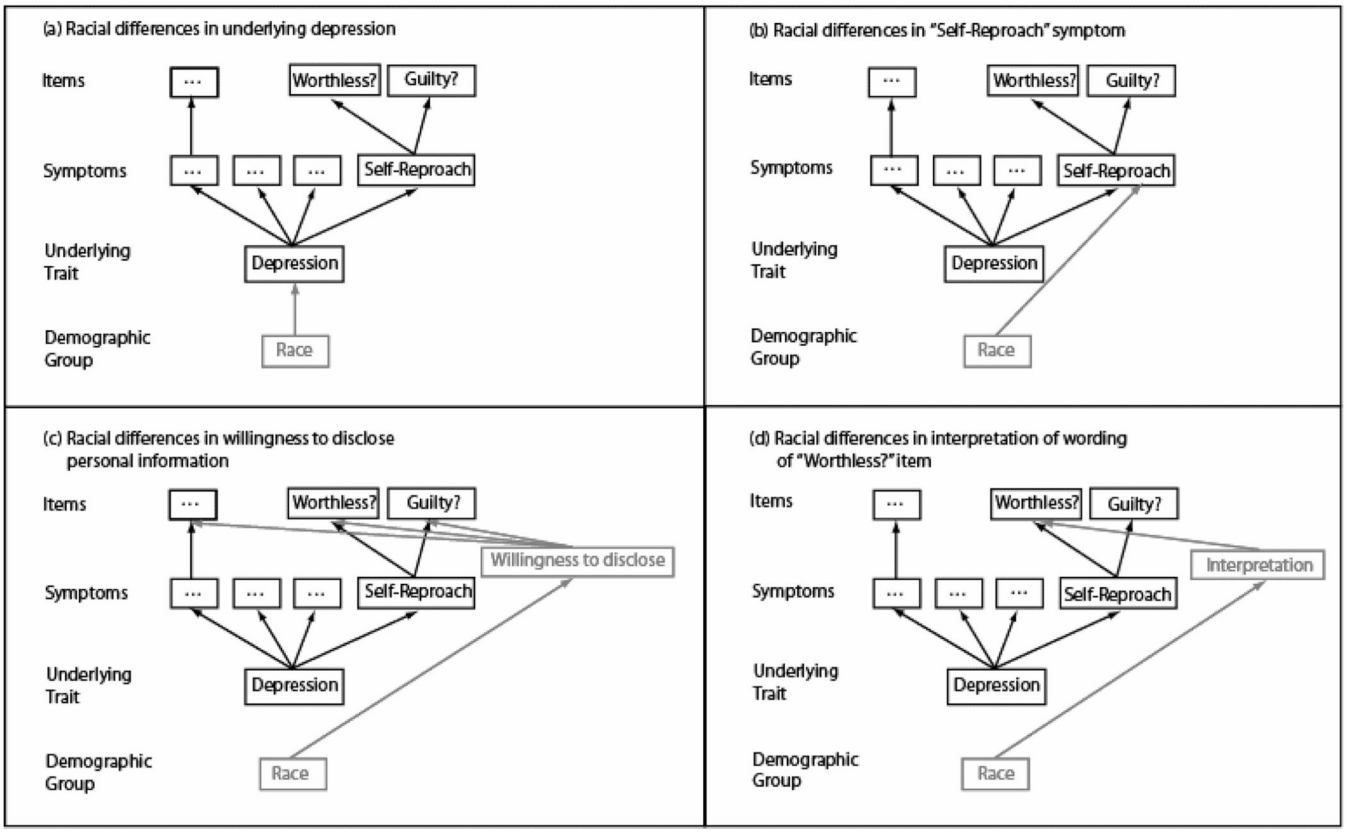

FIGURE 1.
Four possible explanations for group differences in responses to an item asking about feelings of worthlessness.

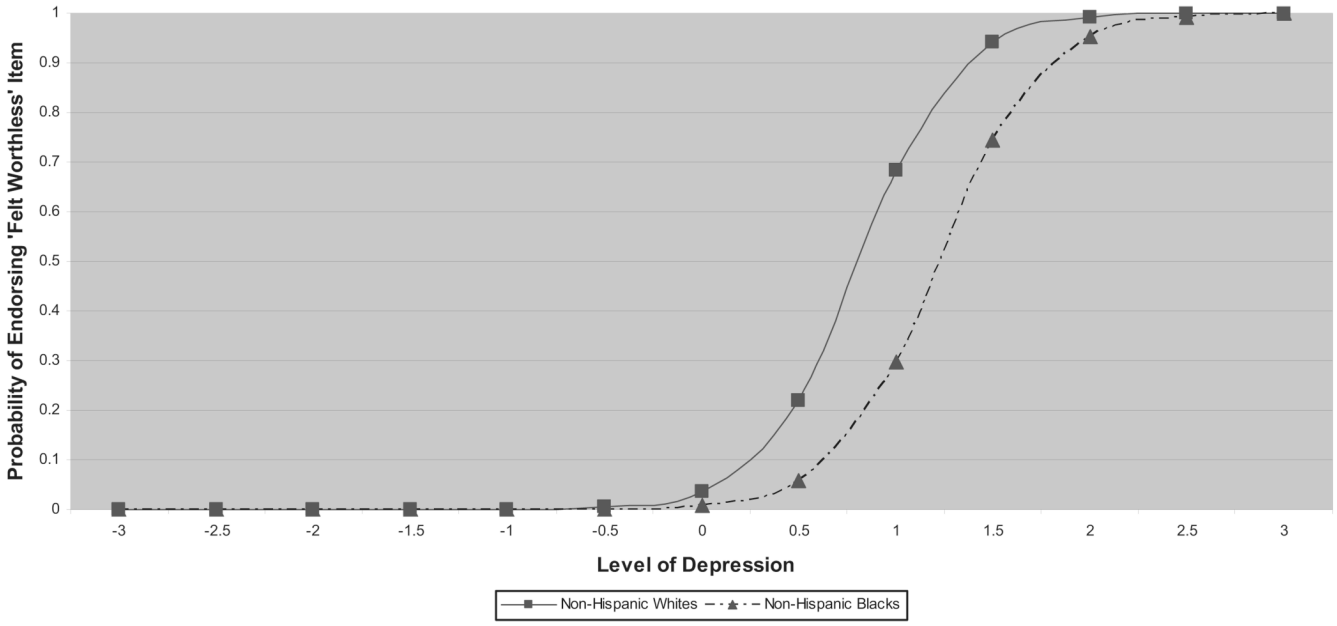

FIGURE 2.
Illustration of differential team functioning (DIF) as variation in item characteristic curves (ICC). ICCs estimated using 2-parameter logistic IRT model.

TABLE 1

Endorsement of Depression Items and Prevalence of Depression Symptoms Across Racial/Ethnic Groups in the NCS[a]

Symptom[b] Item	Full Sample (%)	Whites (%)	Blacks (%)	Hispanics (%)	$\chi^2(2)$	*p*
Anhedonia[c]						
Lost interest	31.2	31.5	26.8	33.2	4.060	0.144
Appetite/weight change	30.5	31.1	**25.9**	31.2	5.424	0.066
Lost appetite	16.2	16.5	**12.2**	18.3	7.234	0.027
Lost weight	15.4	15.5	13.7	16.3	1.360	0.506
Increased appetite	9.9	10.4	**7.0**	9.1	5.650	0.059
Increased weight	7.4	8.0	**4.6**	6.0	8.451	0.015
Sleep disturbance	38.7	39.4	**32.9**	39.4	7.467	0.024
Trouble falling asleep	26.8	26.7	26.9	27.7	0.201	0.904
Trouble staying asleep	21.5	22.0	18.6	20.7	3.057	0.217
Waking early	14.0	14.7	**10.7**	11.3	8.673	0.013
Sleeping too much	11.0	11.0	9.6	12.2	1.683	0.431
Loss of energy[c]						
Lack of energy	29.2	30.3	**21.3**	29.2	16.074	0.000
Retardation/agitation	16.1	16.0	14.3	18.4	2.934	0.231
Slowed down	9.8	9.7	8.7	11.2	1.654	0.438
Couldn't stay still	9.7	9.6	8.2	11.9	3.727	0.155
Self-reproach	24.89	26.28	**15.48**	23.82	26.1083	<0.0001
Felt worthless	17.35	18.2	**9.65**	18.74	21.768	<0.0001
Felt guilty	16.15	16.99	**10.74**	15.2	12.2889	0.0021
Cognitive difficulties	31.28	31.88	28.41	29.55	3.0098	0.222
Trouble concentrating	26.89	27.77	**22.21**	24.88	7.5377	0.0231
Slow thought	13.26	13.56	11.79	12.42	1.4279	0.4897
Trouble deciding	15.29	15.09	15.83	16.32	0.5787	0.7488
Suicidality	32.19	33.77	**26.95**	**25.13**	20.0596	<0.0001
Thoughts of death	26.05	27.04	22.97	**21.39**	9.1243	0.0104
Wanting to die	13.9	14.06	11.43	15.24	3.1345	0.2086
Thoughts of suicide	14.87	16.25	**8.13**	**11.02**	27.3121	<0.0001
Attempted suicide	3.73	3.84	2.38	4.26	2.8536	0.2401
Total sample size (unweighted)	9684	6099	929	784	—	—

Symptom[b] Item	Full Sample (%)	Whites (%)	Blacks (%)	Hispanics (%)	$\chi^2(2)$	*p*
Screen positives(unweighted)	4501	3585	505	411	—	—

[a] Percents are weighted. Statistical tests are design adjusted. Boldface type indicates a signficant chi-square test for difference from whites at the $p = 0.05$ level.

[b] Symptoms were counted as present if any one of the corresponding items were endorsed.

[c] The symptoms anhedonia and loss of energy were each assessed with a single item. The anhedonia item referred to "loss of interest in pleasurable activities" and the loss of energy item referred to a "lack of energy."

TABLE 2

Results of Tests for Differential Item Functioning Against Blacks in Depression Items and Symptoms[a]

		Initial Test for DIF Against Blacks				Final Test for DIF Against Blacks			
		Symptom-Level DIF		Item-Level DIF		Symptom-Level DIF		Item-Level DIF	
Symptom	**Item**	**Est.[b]**	***p*[c]**	**Est.[b]**	***p*[c]**	**Est.[b]**	***p*[c]**	**Est.[b]**	***p*[c]**
Anhedonia	Lost interest	−0.033	0.883	−0.033	0.883	−0.035	0.902	−0.035	0.902
Appetite/weight change	Lost appetite	−0.019	0.749	0.000	0.506	−0.014	0.693	0.015	0.272
	Lost weight			−0.053	0.979			−0.038	0.929
	Increased appetite			0.009	0.327			0.008	0.355
	Increased weight			0.002	0.466			−0.005	0.598
Sleep disturbance	Trouble falling asleep	0.013	0.300	−0.094	1.000	0.019	0.227	−0.080	0.998
	Trouble staying asleep			−0.013	0.684			−0.004	0.559
	Waking early			0.007	0.384			0.017	0.245
	Sleeping too much			0.020	0.183			0.015	0.249
Loss of energy	Lack of energy	**0.076**	**0.002**	**0.076**	**0.002**	**0.085**	**0.000[a]**	**0.085**	**0.000[a]**
Retardation/agitation	Slowed down	−0.024	0.829	−0.029	0.915	−0.020	0.795	−0.028	0.897
	Couldn't stay still			−0.028	0.908			−0.028	0.904
Self-reproach	Felt worthless	**0.093**	**0.000**	**0.064**	**0.002**	0.032	0.079	**Removed**	—
	Felt guilty			0.027	0.119			0.032	0.079
Cognitive difficulties	Trouble concentrating	−0.082	0.999	−0.020	0.768	−0.075	0.998	−0.017	0.734
	Slow thought			−0.052	0.987			−0.039	0.956
	Trouble deciding			−0.083	1.000			−0.074	1.000
Suicidality	Thoughts of death	−0.041	0.928	−0.060	0.983	−0.059	0.983	−0.057	0.980
	Wanting to die			−0.021	0.826			−0.019	0.798
	Thoughts of suicide			**0.077**	**0.000**			**Removed**	—
	Attempted suicide			0.012	0.114			0.009	0.193

Figures in bold type are significant at $p = 0.05$ level using correction for multiple testing (B and H).

[a] Estimates and *p* values from SIBTEST analysis of the NCS data with rough frequency weighting.

[b] Positive estimates indicate DIF against blacks. Negative estimates indicate DIF in favor of blacks.

[c] Significance tests are 1-sided tests for DIF against blacks.

TABLE 3

Results of Tests for Differential Item Functioning Against Hispanics in Depression Items and Symptoms[a]

		Initial Test for DIF Against Hispanics				Final Test for DIF Against Hispanics			
		Symptom Level		Item Level		Symptom Level		Item Level	
Symptom	**Item**	**Est.[b]**	***p*[c]**	**Est.[b]**	***p*[c]**	**Est.[b]**	***p*[c]**	**Est.[b]**	***p*[c]**
Anhedonia	Lost interest	−0.035	0.894	−0.035	0.894	−0.031	0.865	−0.031	0.865
Appetite/weight change	Lost appetite	−0.009	0.634	−0.059	0.986	−0.038	0.919	−0.057	0.984
	Lost weight			−0.021	0.794			−0.020	0.777
	Increased appetite			0.019	0.178			0.019	0.172
	Increased weight			**0.046**	**0.003**[a]			**Removed**	—
Sleep disturbance	Trouble falling asleep	0.012	0.315	0.002	0.471	0.016	0.271	0.008	0.390
	Trouble staying asleep			0.002	0.468			0.006	0.413
	Waking early			**0.071**	**0.001**[a]			**Removed**	—
	Sleeping too much			−0.012	0.689			−0.008	0.635
Loss of energy	Lack of energy	0.037	0.087	0.037	0.087	0.043	0.060	0.043	0.060
Retardation/agitation	Slowed down	−0.029	0.882	−0.014	0.757	−0.020	0.792	−0.005	0.586
	Couldn't stay still			−0.040	0.971			−0.038	0.963
Self-reproach	Felt worthless	0.000	0.504	−0.033	0.910	0.006	0.407	−0.025	0.847
	Felt guilty			−0.025	0.847			−0.021	0.801
Cognitive difficulties	Trouble concentrating	−0.003	0.537	0.013	0.316	0.006	0.418	0.021	0.218
	Slow thought			−0.004	0.568			0.006	0.394
	Trouble deciding			−0.053	0.985			−0.043	0.961
Suicidality	Thoughts of death	**0.078**	**0.003**[a]	0.043	0.054	0.049	0.037	0.046	0.044
	Wanting to die			−0.043	0.969			−0.041	0.962
	Thoughts of suicide			**0.052**	**0.010**[a]			**Removed**	—
	Attempted suicide			0.001	0.476			0.001	0.463

Figures in bold type are significant at $p = 0.05$ level using correction for multiple testing (B and H).

[a] Estimates and *p* values from SIBTEST analysis of the NCS data with rough frequency weighting.

[b] Positive estimates indicate DIF against Hispanics. Negative estimates indicate DIF in favor of Hispanics.

[c] Significance tests are 1-sided tests for DIF against Hispanics.