Published in final edited form as:

J Clin Epidemiol. 2007 November ; 60(11): 1195–1200.

Center for Epidemiologic Studies—Depression Scale (CES-D) Item Response Bias Found with Mantel-Haenszel Method Successfully Replicated Using Latent Variable Modeling

Frances M. Yang^{1,2,*§} and Richard N. Jones^{1,3,*}

¹*Institute for Aging Research, Hebrew SeniorLife, Boston, Massachusetts*

²*Department of Psychiatry, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts*

³*Division of Gerontology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts*

Abstract

Objective—This study reexamines findings reported by Cole and colleagues of item-response bias in the Center for Epidemiologic Studies—Depression (CES-D) scale by age, gender, and race. We use an item response theory (IRT) based latent variable conditioning approach.

Study Design and Setting—We used the multiple indicators, multiple causes model (MIMIC) model framework to estimate measurement bias in the CES-D responses of participants in the New Haven Established Populations for the Epidemiologic Studies of the Elderly (EPESE) study (N=2,340).

Results—Measurement bias attributable to race was significant for the following two CES-D items: people “are unfriendly” and “dislike me.” The proportional odds of a higher category response by blacks relative to whites on these items were 2.35 (95% confidence interval [CI]: 1.65, 3.36) and 3.11 (95% CI: 2.04, 4.76), respectively. The proportional odds were higher among women (2.03 [95% CI: 1.35, 3.06]) relative to men for the CES-D item “crying.”

Conclusion—Our findings confirm three items on the CES-D show strong evidence of item-response bias. The MIMIC model is preferable to the Mantel-Haenszel approach because it conditions on a latent variable, while the effect estimates can also be interpreted using a proportional odds framework.

Keywords

Center for Epidemiological Studies—Depression (CES-D); late-life depression; multiple indicators; multiple causes (MIMIC) model; Established Populations for the Epidemiologic Studies of the Elderly (EPESE); differential item functioning (DIF); item response theory (IRT)

'What is new?' Text Box

§Corresponding author E-mail addresses: FMY: francesyang@hrca.harvard.edu, RNJ: jones@hrca.harvard.edu.

*These authors contributed equally to this work

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

- Center for Epidemiologic Studies—Depression Scale (CES-D) Item Response Bias found with Mantel-Haenszel Method was successfully replicated using Latent Variable Modeling
- Expression of depression in older adults is influenced by race/ethnicity and sex. Three items on the CES-D show strong evidence of item-response bias: people "are unfriendly", "dislike me" and 'crying'
- The multiple indicators, multiple causes (MIMIC) model. MIMIC model can produce results that are interpretable to a public health research audience (i.e., proportional odds ratios expressing DIF)
- The methodology is not very complicated (but does require special software) and most importantly, is theoretically appropriate and inferentially superior.

1. Introduction

Detecting item response bias is important to avoid compromising the validity of an entire scale. Item response bias occurs when one or enough items in the scale exhibit differential item functioning (DIF), or measurement bias (measurement noninvariance), due to group membership. More than five years ago, Cole and colleagues (1) examined the Center for Epidemiologic Studies-Depression (CES-D) scale used in the New Haven Established Populations for the Epidemiologic Studies of the Elderly (EPESE) for item bias related to age, gender, and race. The methodology used in their study was an extension of the Mantel-Haenszel (MH) adjustment (2), which the authors argued was "most appropriate for a medical and public health audience due to the use of proportional odds ratios" (p.286) (1).

Cole and colleagues (1) produced one of the first studies examining the effects of sociocultural characteristics on the measurement properties of the CES-D among older adults. There have been multiple studies that have used other methods developed from item response theory to examine other depression measures for the presence of DIF attributable to different sociocultural characteristics (3,4). Some of these approaches, in addition to the proportional odds approach used in Cole et al. (1), have included Thissen's (5) item response theory likelihood-ratio tests for DIF; Raju's (6) differential functioning of item and tests; and the multiple indicators, multiple causes (MIMIC) model (7). Within the structural equation modeling framework, the MIMIC model can be viewed as a special case of a confirmatory factor analytic model for dichotomous dependent variables with covariates, as described in Jones and Gallo (8). In a simulation study comparing the MIMIC model, MH, item response theory likelihood-ratio, and simultaneous item bias test, the MIMIC model was the most accurate method for detecting DIF (9).

This paper focuses on the MIMIC model, which has been used in a number of health-related studies (8,10–17), and uses the same data and addresses the same questions as Cole and colleagues (1). We reexamined the presence of measurement noninvariance in CES-D items that may be attributable to age, gender, and race. We hypothesized that the MHOR model, which is the observed-score conditioning approach used by Cole and colleagues (1), will identify more CES-D items with DIF, than the MIMIC method.

2. Methods

2.1. Study Population

We obtained the New Haven Established Populations for Epidemiologic Studies of the Elderly (EPESE) public use data via the Interuniversity Consortium for Political and Social Research. New Haven, Connecticut (USA), was one of four EPESE sites contracted by the National

Institute on Aging (NIA) to conduct a population-based survey of adults aged 65 and older to provide information about risks for mortality, hospitalization, and placement in long-term care facilities and to investigate risk factors for chronic diseases and loss of functioning. Details of the EPESE are presented elsewhere (18). Specifically, the collection of data at the New Haven site began in 1981 and the first wave was completed in 1982. These are the data used for this study. Just as in Cole and colleagues (1), exclusion criteria included a missed response on at least one CES-D question or background variable (age, gender, or race) (N=472). The final comparable sample used for this study included respondents who provided complete data on all CES-D questions and background variables (N=2,340).

It is worth pointing out that our analytic approach (*Mplus* MIMIC modeling) can provide parameter estimates in the case of missing dependent variables under missing at random assumptions. We have also included a larger sample of 2,733 older individuals using the cases that included a missed response on at least one, but not all, CES-D question (N=393), assuming the data that are missing are missing at random.

2.2. Measurements

2.2.1. Depression assessment—Each EPESE site used a variation of the CES-D (19) to assess depression symptoms. The New Haven respondents were presented with the original version of the 20 CES-D items with a four-category response (rarely or none of the time, some of the time, much of the time, most or all of the time).

2.2.2. Background variables—Age is available in the public use data files as a discrete categorical variable. In our models, age was treated as a binary variable (age <75=0, ≥75=1). Gender was also a binary variable (male=0, female=1). The two categories for race were Black (=1) and White (=0).

2.2.3. Statistical Analyses—The analytic approach involved three main stages conducted in the *Mplus* software, version 4.1 (7). The first stage involved estimating a MIMIC model without direct effects: essentially a confirmatory factor analysis model with covariates that influence a single underlying latent trait. This model includes only the so-called indirect effects in a single group MIMIC model (20), and as such do not include any DIF effects. Indirect effects are the relationships between the covariates and observed dependent variables (latent factor indicators) mediated by latent factors, which are simultaneously estimated in the MIMIC model (21). The modeling framework for this and the following stage was multivariate probit using the *Mplus* weighted least squares mean and variance adjusted estimator (7).

The second stage was a forward stepwise model building procedure to identify significant direct effects, which reflect the presence of uniform DIF. We obtained the matrix of fit derivatives (scaled as chi-square and referred to as modification indices) for the regressions of CES-D items on age, gender, and race. The direct effects are relationships between the covariates and latent factor indicators that are not mediated by the latent factors themselves (22). The modification indices reveal, where possible, the direct effects of background variables on the CES-D items that, if freely estimated, would significantly improve model fit.

To evaluate the significance of model modifications implied by the modification indices, we performed a robust chi-square model difference testing using the *Mplus* DIFFTEST function (suitable for testing nested models with the WLSMV estimator). The DIFFTEST function compares the fit of a null model (H_0 : MIMIC model with no direct effects) to an alternative model (H_1 : MIMIC model with at least one direct effect). We first compared the H_0 model to H_{1a} (MIMIC model with one direct effect) and then compared H_{0a} (which was H_{1a}) to H_{1b} (MIMIC model with two direct effects). The test of chi-square difference between each null

and alternative model continued until the final fitted model, with all significant direct effects freed, was no longer significant using an alpha level of 0.05.

Lastly, we obtained estimates of bias (direct effects) on a proportional odds scale by re-estimating our final fitted model using the *Mplus* MLR estimator, which implements a multivariate logistic regression framework and taking the exponentiation of the logistic regression coefficients (odds ratios). We then determined if there was a “relatively large” meaningful measurement bias in each item that was attributable to any of the background variables by considering the following rule of Cole and colleagues (1): proportional odds ratio >2.0 or <0.5 shows that those in the test group have more than twice the “odds of responding higher to the individual item than those in the control group, after being matched on overall depressive symptoms.”

3. Results

Sample characteristics and the CES-D item responses for the 2,340 participants can be found in Cole and colleagues (1). The DIFFTEST results are as follows: The chi-square difference between the first H_0 model (MIMIC model with no direct effects) and the alternative model (H_{1a} : MIMIC model with the direct effect of “people dislike me” and race) was significant ($\chi^2=38.55$, $df=1$, $p<0.001$). Therefore, we continued to test subsequent null and alternative models until we reached the final nested models with all significant direct effects freed. The chi-square difference between the second H_0 model (H_{0a} : MIMIC model with the direct effect of “people dislike me” on race) and the alternative model (H_{1b} : MIMIC model with the direct effects of “people dislike me” and “people are unfriendly” on race) was significant ($\chi^2=33.26$, $df=1$, $p<0.001$). The chi-square difference between the final H_{0b} model (MIMIC model with the direct effects of “people dislike me” and “people are unfriendly,” on race) and the H_{1c} (MIMIC model with the direct effect of “people dislike me,” “people are unfriendly,” and “crying spells” on race) was still significant ($\chi^2=22.23$, $df=1$, $p<0.001$).

The final MIMIC model showed measurement noninvariance in three CES-D items attributable to race and gender. The proportional odds of blacks responding in a higher category than whites on two CES-D items, “people are unfriendly” and “people dislike me” were 2.35 (95% confidence interval [CI]: 1.65, 3.36) and 3.11 (95% CI: 2.04, 4.76), respectively. The proportional odds ratio was higher for women [2.03 (95% CI: 1.35, 3.06)] than for men for endorsing the CES-D item “crying spells.”

We also estimated MIMIC models for 2,733 older adults, the sample resulting from including cases with a missing response on at least one, but not all, CES-D question ($N=393$) (Table 1). This is one of the advantages of the *Mplus* approach. Results showed measurement noninvariance in two CES-D items attributable to race. The proportional odds of blacks responding in a higher category than whites on two CES-D items, “people are unfriendly” and “people dislike me” were 2.30 (95% confidence interval [CI]: 1.78, 2.99) and 3.08 (95% CI: 2.26, 4.19), respectively. The proportional odds ratio was higher for women [1.91 (95% CI: 1.43, 2.56)] than for men for endorsing the CES-D item “crying spells,” but the effect size did not reach the threshold for clinical significance as stated by Cole and colleagues (1).

4. Discussion

The final model using the 2,340 comparable subjects (excluding participants with missing data) confirmed the main results of Cole and colleagues (1): 17 of the 20 CES-D items were observed to be relatively free of DIF attributable to age, gender, and race. The three items found to have DIF were identical to those of Cole and colleagues (1). There are race differences in endorsing interpersonal depressive symptoms; older blacks are more likely than older whites to say that

“people are unfriendly” and that “people dislike me.” As for gender differences, older women were more likely than older men to admit that they experienced “crying spells.”

Some authors argue for dropping items with DIF from a scale. For example, Cole and colleagues (1) found that the 17-item CES-D “retained a high internal consistency reliability” (p.288). This is not surprising given that the item set that includes items with DIF is less unidimensional than an item set that excludes items with DIF, following the conceptualization of DIF as a form of multidimensionality as articulated by Oort (23). However, it does not necessarily follow that the reduced item set will be more or less valid for a specific use. Therefore, one of the advantages of the MIMIC approach is that this measurement noninvariance in CES-D responses can be brought into the model and statistically controlled when investigating other substantively focused research questions that may involve causal or confounded effects of participant background variables. The significance of examining DIF due to sociocultural differences in the CES-D is to ensure that older adults with the same underlying depressive trait from different gender or race groups receive equivalent scores.

The technique of Cole and colleagues, an extension of the MH approach, can be viewed as a contribution to the field of psychometrics by that of epidemiology (24). This nonparametric method is based on *observed total test score conditioning* (25). While controlling for the level of depression, if the conditional odds of endorsing the test item are the same for both groups (i.e., those ≥ 75 vs. those < 75 years of age), then DIF is not present. However, if there is a significant interaction between the item and the group, then DIF is present. Relevant to this study, the advantage of the MHOR approach is that there are few model assumptions, relatively simple statistical procedures, and results that are easily interpretable for the health-research audience. One major disadvantage of this model is that modeling multiple characteristics of the respondents that may be attributable to DIF simultaneously is more complex and difficult than examining one characteristic at a time.

Under the MIMIC model, we were able to examine multiple background variables and used an *unobserved or latent conditioning variable*, which is the underlying depressive trait for individuals. The parametric method is based on *unobserved or latent test score conditioning* (26). There is a very strong conceptual advantage for detecting measurement bias using a latent or unobserved variable to condition the likelihood. When conditioning on the total score (an observed score conditioning approach), it is a fact that a average item-test regression line is a straight line (27). Therefore, when looking at individual items, any item with DIF indicating disadvantage for a focal group must be met with evidence of DIF in other items indicating disadvantage for the reference group. In real data situations, differentiating true DIF from statistical artifact is not possible. Thus, using the MH approach can lead to cases of Berkson's bias and other interpretative difficulties in DIF detection methods (28). As pointed out by Meredith and Millsap (29): “item bias detection procedures that rely strictly on observed variables are not in general diagnostic of measurement bias, or the lack of measurement bias.”

It is worth pointing out that Cole and colleagues used a log-transformed latent trait estimate. Therefore, the straight-line average item-test regression limitation of MH approaches may not hold. In fact, early IRT algorithms (e.g., Lord's LOGIST procedure) used the log of the number right over number wrong as an initial estimate of the latent trait. However, these mathematical manipulations are close to, but not the same as, a latent variable and full-IRT based method for detecting DIF, as we have demonstrated in this manuscript.

In the current case, although the two methods rely on different assumptions, the results are comparable because the single group MIMIC model and the MHOR approach both rely upon a proportional likelihood model (8). That is, group differences in the likelihood of endorsing

a particular symptom are viewed as proportional over the range of the conditioning variable. In the MHOR approach used by Cole and colleagues (1), the conditioning variable was a transformed raw score. In the MIMIC approach, the conditioning variable is a latent variable, analogous to a factor in confirmatory factor analysis. Moreover, the MIMIC approach we use is model-based, iterative, and arrives at a final parsimonious model that accounts for specific patterns of symptom endorsement for specific groups, which are presumed to reflect measurement bias. We obtained proportional odds by re-estimating our final fitted model (obtained using multivariate probit regression and estimated with a WLSMV estimator) using a MLR-based multivariate logistic regression approach (MLR/LR) (7).

It is worth mentioning that the software used to estimate the models in this study is, relative to the MH approach, a very recent addition to the potential armamentarium of applied researchers. The *Mplus* MLR logistic regression was introduced in 2004, while the MHOR approach was introduced in 1959 (30). We expect that as more researchers become aware of this approach, it will be extended into many different substantive areas in public health research.

5. Conclusions

By using the single group MIMIC model, we obtained similar numeric results interpreted in the same manner (as proportional odds) as did Cole and colleagues (1) using a MH proportional odds approach. With regard to the extension of the MH as “most appropriate for a medical and public health audience due to the use of proportional odds ratio” (1), we believe that caution needs to be exercised in observed score conditioning approaches given the previous reports in the psychometric literature expressing the severe inferential limitations of such approaches (28).

There are many advantages to the latent variable conditioning approach that we have illustrated as implemented with *Mplus* software. These include the ability to control for putative bias in more than one background variable at a time, to model effects of background variables that are continuously distributed, and the to handle missing item-level data under missing at random assumptions. In addition, the *Mplus* software is capable of incorporating sampling weights and complex sample designs, as well as, testing and potentially controlling for violations of the proportionality assumption (i.e., non-uniform differential item functioning). Most importantly, measurement models with parameters capturing noninvariance effects can be incorporated into larger and substantively meaningful models (e.g., survival models, risk models); and therefore, introducing statistical adjustment for putative bias effects without dropping assessment items.

Acknowledgement

This research was made possible through the National Institutes of Health (NIH)/National Institute on Aging (NIA) 5-T32 AG023480 award and NIA grants AG025308, AG008812, and AG021153.

References

1. Cole SR, Kawachi I, Maller SR, Berkman LF. Test of item-response bias in the CES-D scale: Experience from the New Haven EPESE Study. *Journal of Clinical Epidemiology* 2000;53:285–289. [PubMed: 10760639]
2. Holland, P.; Wainer, H. *Differential Item Functioning*. New York: Lawrence Erlbaum Associates; 1993.
3. Millsap RE, Everson HT. Methodology review: Statistical approaches assessing measurement bias. *Applied Psychological Measurement* 1993;17:297–334.
4. Hambleton, RK.; Swaminathan, H.; Rogers, HJ. *Fundamentals of Item Response Theory*. Newbury Park, CA: SAGE Publications, Inc; 1991. Identification of Potentially Biased Test Items.

5. Thissen, D. IRTLRDIF v.2.0b: Software for the Computation of the Statistics Involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning. 2.0b ed.. University of North Carolina at Chapel Hill: L.L. Thurstone Psychometric Laboratory; 2001.
6. Raju NS, van der Linden WJ, Fleer PF. IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement* 1995;19:353–368.
7. Muthén, LK.; Muthén, BO. Mplus Version 4.1. Los Angeles: Muthén & Muthén; 1998–2006.
8. Jones RN, Gallo JJ. Education and sex differences in the Mini Mental State Examination: Effects of differential item functioning. *The Journals of Gerontology Series B, Psychological Sciences and Social Sciences* 2002;57B(6):P548–P558.
9. Finch H. The MIMIC Model as a Method for Detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT Likelihood Ratio. *Applied Psychological Measurement* 2005;29(4):278–295.
10. Jones R, Gallo J. Education bias in the Mini-Mental State Examination. *International Psychogeriatrics* 2001;13(3):299–310. [PubMed: 11768377]
11. Jones RN. Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination: Detecting differential item functioning using MIMIC modeling. *Medical Care* 2006;44(11 Suppl 3):S124–S133. [PubMed: 17060819]
12. Jones RN. Racial bias in the assessment of cognitive functioning of older adults. *Aging & Mental Health* 2003;7(2):83–102. [PubMed: 12745387]
13. Van de Ven WP, Van der Gaag J. Health as an unobservable: a MIMIC-model of demand for health care. *Journal of Health Economics* 1982;1(2):157–183. [PubMed: 10263954]
14. Gallo JJ, Rabins PV, Anthony JC. Sadness in older persons: 13-year follow-up of a community sample in Baltimore, Maryland. *Psychol Med* 1999;29(2):341–350. [PubMed: 10218925]
15. Bogner HR, Gallo J. J. Are higher rates of depression in women accounted for by differential symptom reporting? *Social Psychiatry and Psychiatric Epidemiology* 2004;39(2):126–132. [PubMed: 15052394]
16. Mast BT, MacNeill SE, Lichtenberg PA. A MIMIC model approach to research in geriatric neuropsychology: The case of vascular dementia. *Aging, Neuropsychology, and Cognition* 2002;9(1):21–37.
17. Mast BT. Cerebrovascular disease and late-life depression: a latent-variable analysis of depressive symptoms after stroke. *American Journal of Geriatric Psychiatry* 2004;12(3):315–322. [PubMed: 15126233]
18. Taylor, JO.; Wallace, RB.; Ostfeld, AM.; Blazer, DG. 1981–1993: (East Boston, Massachusetts, Iowa and Washington Counties, Iowa, New Haven, Connecticut, and North Central North Carolina). [Computer File]. 3rd ICPSR version. Computer File. Bethesda, MD: National Institute on Aging (producer), 1997. Ann Arbor, Michigan: Interuniversity Consortium for Political and Social Research (distributor), 1998. 1998. Established populations for epidemiologic studies of the elderly.
19. Radloff L. The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement* 1977;1:385–401.
20. Jöreskog K, Goldberger A. Estimation of a model of multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association* 1975;10:631–639.
21. Muthén BO. Latent variable modeling in heterogeneous populations. *Psychometrika* 1989;54(4):557–585.
22. Muthén, B.; Muthén, LK. Mplus Short Courses. 2005 [cited 2005 November 7]. Traditional Latnet Variable Modeling Using Mplus; p. 161 Available from:
23. Oort F. Using restricted factor analysis to detect item bias. *Methodika* 1992;6:150–166.
24. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 1959;22:719–748. [PubMed: 13655060]
25. Teresi JA, Golden R. Latent structure methods for estimating item bias, item validity and prevalence using cognitive and other geriatric screening measures. *Alzheimer Dis Assoc Disord* 1994;8:S291–S298. [PubMed: 8068271]
26. Teresi J, Golden R. Latent structure methods for estimating item bias, item validity and prevalence using cognitive and other geriatric screening measures. *Alzheimer Disease and Associated Disorders* 1994;8:S291–S298. [PubMed: 8068271]

27. Lord, F.; Novick, M. Statistical theories of mental test scores. Reading, MA: Addison-Wesley; 1968.
28. Camilli, G. The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues. In: Holland, P.; Wainer, H., editors. Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers; 1993. p. 397-413.
29. Meredith W, Millsap R. On the misuse of manifest variables in the detection of measurement bias. *Psychometrika* 1992;57(2):289–311.
30. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 1959;22:719–748. [PubMed: 13655060]

Table 1

Results of differential item functioning detection using Mantel-Haenszel Odds Ratio Approach (Cole et al., 2000) and using MIMIC model: proportional odds ratios between CES-D items and age, gender, and race, conditioned on total CES-D score^a (Cole et al., 2000) or latent variable (MIMIC)^b in New Haven EPESE study.

Item	Proportional odds regression model (N=2340) ^e	Age (75 and older) MIMIC (N=2340) <i>de</i>	MIMIC (N=2733) <i>df</i>	Proportional odds regression model (N=2340) ^e	Sex (Female) MIMIC (N=2340) <i>de</i>	MIMIC (N=2733) <i>df</i>	Proportional odds regression model (N=2340) ^e	Race/ethnicity (Black) MIMIC (N=2340) <i>de</i>	MIMIC (N=2733) <i>df</i>
1) Bothered by things	1.13	1.00	1.00	1.01	1.00	1.00	0.68	1.00	1.00
2) Poor appetite	0.94	1.00	1.22	0.93	1.00	1.00	1.40	1.00	1.67
3) Could not shake the blues	1.09	1.00	1.00	1.25	1.00	1.00	0.79	1.00	1.00
4) As good as others	1.01	1.00	1.00	1.04	1.00	1.00	0.91	1.00	1.00
5) Trouble sleeping	1.38	1.00	1.33	0.98	1.00	1.00	0.83	1.00	1.00
6) Felt depressed	0.87	1.00	1.00	1.08	1.00	1.00	0.77	1.00	1.00
7) Everything an effort	1.10	1.00	1.31	0.89	1.00	0.85	1.04	1.00	1.35
8) Felt hopeless	1.13	1.00	1.31	0.84	1.00	0.86	1.06	1.00	1.30
9) Life a failure	0.71	1.00	0.73	0.70	1.00	0.68	0.99	1.00	1.00
10) Felt fearful	0.97	1.00	1.00	1.35	1.00	1.00	1.12	1.00	1.00
11) Restless sleep	0.91	1.00	1.00	1.28	1.00	1.00	0.85	1.00	1.00
12) Happy	0.91	1.00	1.00	1.14	1.00	1.00	0.77	1.00	1.00
13) Talked less	0.99	1.00	1.00	0.70	1.00	1.00	0.99	1.00	1.00
14) Felt lonely	1.15	1.00	1.22	1.35	1.00	1.00	0.96	1.00	1.00
15) People unfriendly	0.74	1.00	1.00	0.74	1.00	0.69	2.29 ^c	2.35	2.30
16) Enjoyed life	1.12	1.00	1.00	1.07	1.00	1.00	0.68	1.00	1.00
17) Crying spells	1.26	1.00	1.00	2.14 ^c	2.03	1.91	0.62	1.00	0.62
18) Felt sad	0.89	1.00	1.00	1.30	1.00	1.00	0.93	1.00	1.00
19) People disliked me	0.73	1.00	1.00	0.85	1.00	0.73	2.96 ^c	3.11	3.08
20) Could not get going	0.87	1.00	1.00	1.24	1.00	1.00	1.04	1.00	1.00

^a Log transformed CES-D score (Cole et al., 2000)

^b Simultaneously controlling for all age, sex, and race/ethnicity covariates with purified trait

^c Meets criteria for proportional odds ratio >2.0 or <.5 demonstrating a "relatively large" meaningful bias

^d MIMIC direct effect estimates scaled as ORs that are exactly 1.00 imply direct effect estimates not estimated in the final fitted model

^e Listwise complete sample

^f Sample with missing data on some, but not all CES-D items