Title

Differential item functioning due to cognitive status does not impact depressive symptom measures in four heterogeneous samples of older adults

Permalink
https://escholarship.org/uc/item/9x80449q

Journal
International Journal of Geriatric Psychiatry, 30(9)

ISSN
0885-6230

Authors
Fieo, Robert
Mukherjee, Shubhabrata
Dmitrieva, Natalia O
et al.

Publication Date
2015-09-01

DOI
10.1002/gps.4234

Peer reviewed

Published in final edited form as:
Int J Geriatr Psychiatry. 2015 September ; 30(9): 911–918. doi:10.1002/gps.4234.

Differential item functioning due to cognitive status does not impact depressive symptom measures in four heterogeneous samples of older adults

Robert Fieo[1], **Shubhabrata Mukherjee**[2], **Natalia O. Dmitrieva**[3], **Denise C. Fyffe**[4], **Alden L. Gross**[5,6], **Elizabeth R. Sanders**[2], **Heather R. Romero**[7,8], **Guy G. Potter**[3,7], **Jennifer J. Manly**[1], **Dan M. Mungas**[9], and **Laura E. Gibbons**[2]

[1]Cognitive Neuroscience Division, Department of Neurology and Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University, New York, NY, USA

[2]General Internal Medicine, University of Washington, Seattle, WA, USA

[3]Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Durham, NC, USA

[4]Kessler Foundation, Spinal Cord Injury/Outcomes and Assessment Laboratory and New Jersey Medical School, Rutgers University, West Orange, NJ, USA

[5]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

[6]Center on Aging and Health, Johns Hopkins University, Baltimore, MD, USA

[7]Joseph and Kathleen Bryan Alzheimer's Disease Research Center, Duke University Medical Center, Durham, NC, USA

[8]Department of Psychology, University of Notre Dame, Notre Dame, IN, USA

[9]Department of Neurology, University of California, Davis Medical Center, Sacramento, CA, USA

Abstract

Objective—The objective of this study is to determine whether differential item functioning (DIF) due to cognitive status impacted three depressive symptoms measures commonly used with older adults.

Methods—Differential item functioning in depressive symptoms was assessed among participants (N = 3558) taking part in four longitudinal studies of cognitive aging, using the Geriatric Depression Scale, the Montgomery–Åsberg Depression Rating Scale, and the Center for Epidemiologic Studies Depression Scale. Participants were grouped by cognitive status using a general cognitive performance score derived from each study's neuropsychological battery and linked to a national average using a population-based survey representative of the US population.

Correspondence to: R. Fieo, raf2168@cumc.columbia.edu.

Conflict of interest
None declared.

The Clinical Dementia Rating score was used as an alternate grouping variable in three of the studies.

Results—Although statistically significant DIF based on cognitive status was found for some depressive symptom items (e.g., items related to memory complaints, appetite loss, lack of energy, and mood), the effect of item bias on the total score for each scale was negligible.

Conclusions—The depressive symptoms scales in these four studies measured depression in the same way, regardless of cognitive status. This may reduce concerns about using these depression measures in cognitive aging research, as relationships between depression and cognitive decline are unlikely to have been due to item bias, at least in the ways that were measured in the datasets we considered.

Keywords

depressive symptoms; differential item function; cognitive impairment; item bias

Introduction

There is mounting evidence to support the functional relationship between cognitive impairment and depression in older adults. A review reported that depression was one of only a few factors to show a consistent association with Alzheimer's disease or cognitive decline across multiple studies (Williams *et al*, 2010). Proposed explanations for the association include (i) depression is an early prodrome of dementia, (ii) depression is a clinical manifestation of dementing diseases, and (iii) depression leads to damage to the hippocampus by way of a glucocorticoid cascade (Jorm, 2001). To study these hypotheses effectively requires a depressive symptoms instrument with measurement equivalence in both healthy and cognitively impaired people.

Too often, measurement equivalence has been assumed in research on depression and cognitive impairment, without supportive data. Fortunately, it can be tested directly. One approach is by testing for differential item functioning (DIF). DIF assesses whether certain test items perform differently for ex-aminees from one group compared with another; test items may measure "different constructs" for those with cognitive impairment than for those without impairment. It may be, for example, that among those with cognitive impairment, questions related to memory are related to the cognitive impairment rather than just to depression. We are aware of only one study that examined DIF in depression for groups differing by cognitive ability. In a sample of Hong Kong Chinese patients with lung disease, Tang *et al* (2005) found that no items demonstrated significant DIF for cognitive impairment. However, the authors convey that the lack of DIF was due to only 3% of the sample presenting with a Mini-mental state examination below 15 and mention that McGivney *et al* (1994) indicate that the Geriatric Depression Scale (GDS) will remain valid until this 15-point threshold. Another study found that the GDS (Sheikh and Yesavage, 1986) had reduced sensitivity for a diagnosis of major depressive disorder in Alzheimer's patients, compared with those with normal cognition (Gilley and Wilson, 1997). Both findings were based on small sample sizes that were *not* broadly generalizable to racially/ethnically diverse groups of healthy and non-healthy persons.

The aim of this paper is to establish whether systematic error in the form of item bias has compromised our ability to measure depression in older adults accurately. We examined three commonly used depressive symptoms measures, the GDS (Sheikh and Yesavage, 1986), the Montgomery–Åsberg Depression Rating Scale (MADRS; Montgomery and Asberg, 1979), and the Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977) in four large, diverse studies of the older people in the USA. We hypothesized that some depressive symptoms reported by older adults would be biased by cognitive ability; DIF analysis would identify a few scale items that, if not addressed, could potentially interfere with the accuracy of the instrument used to measure depression.

Methods

Participants

Differential item functioning in depressive symptoms was assessed among participants (*N* = 3558) taking part in the following four longitudinal studies of cognitive aging: the Joseph and Kathleen Bryan Alzheimer's Disease Research Center at Duke University (*Duke ADRC*; *N* = 511; Carvalho *et al*, 2015), the University of California, Davis' Alzheimer's Disease Center Longitudinal Cohort (*UC Davis ADC*; *n* = 620; Hinton *et al*, 2010), the Washington Heights/Hamilton Heights-Inwood Columbia Aging Project (*WHICAP*; *n* = 2,137; Manly *et al*, 2005), and the Neurocognitive Outcomes of Depression in the El-derly (*NCODE*) Study (*n* = 290; Steffens *et al*, 2004). We examined data at the first occasion when depressive symptoms were assessed, usually at baseline. The four samples had notable demographic heterogeneity (Table 1), but the depression measures were not different because of age, sex, years of education, race, ethnicity, or test language (Dmitrieva *et al*, 2014). The *Duke ADRC*, *UC Davis ADC*, and *NCODE* studies did not exclude people with dementia at baseline; *WHICAP* excluded people if they or their caregivers reported that they had significant cognitive problems. Clinically significant depression symptoms were found in 21.5% of subjects (Grunebaum *et al*, 2008). NCODE enrolled individuals with a diagnosis of major depressive disorder (70% of the sample) and controls without depression (30%). The other studies did not conduct depression diagnoses; however, we report the percentage of each sample scoring above the conventional cut off score for the depression measure used.

Measures

Differential item functioning in depressive symptoms was examined across the GDS, the CES-D, and the MADRS. The GDS is a 15-item scale, with dichotomous response options, that is a well-validated and highly reliable scale among the older adult population (for review, see Montorio and Izal, 1996). It was administered in the *Duke ADRC* and *UC Davis ADC*. Of the three scales, the 20-item version of the CES-D assesses the highest number of depressive symptoms, using multiple response categories. This version was administered in *NCODE*. *WHICAP* administered the 10-item version, using a dichotomous response option. Both versions of the CES-D have been validated for use among older adults (Andresen *et al*, 1994; Hertzog *et al*, 1990). *NCODE* also administered the MADRS, a 10-item scale that has also been validated among older adult populations (Engedal *et al*, 2012; Mottram *et al*, 2000).

Cognitive status grouping variables

To explore DIF between different levels of cognitive status, we grouped individuals on the basis of three different criteria, depending on the data available in each study.

1. General cognitive performance (GCP). Items from each study's neuropsychological battery were linked to results from a population-based survey that is representative of older adults in the USA, using calibration methods previously reported in Gross *et al* (2014a, 2014b). For this investigation, these scores were dichotomized into categories of above and below the mean national score.
2. Clinical Dementia Rating (CDR) score (Morris, 1993). In *Duke ADRC, UC Davis ADC*, and *WHICAP*, those with normal CDR (CDR = 0) were compared with those with questionable, mild, moderate, and severe dementia (CDR = 0.5–3). In *WHICAP*, there were enough participants to also compare CDR scores of 0 and 0.5 to 1–3.
3. Clinical consensus diagnosis. *UC Davis ADC* participants also had a clinical consensus diagnosis, which was collapsed to compare "no cognitive impairment" with the categories ranging from "questionable impairment" to "dementia" (Mungas *et al*, 2010).

Statistical analyses

Differential item functioning was present if different groups of individuals (e.g., impaired versus non-impaired or higher versus lower general cognition) at similar levels of depression exhibited different probabilities of individual item scores (Hulin, 1987). The literature on DIF detection is diverse, and there are a multitude of methods available, such as contingency table, item response theory (IRT), structural equation modeling, and logistic regression methods (Scott *et al*, 2010). All of these methods seek to identify which items function the same and which items need to be estimated separately in each group, by allowing the item parameters to vary across groups and testing whether these differences are statistically significant. We employed a logistic regression/IRT approach, using the lordif software available in the Comprehensive R Archive Network (Choi *et al*, 2011).

The first stage of the DIF analysis began by using IRT to estimate the underlying level of depression. Samejima's Graded Response Model was used for ordinal variables (Samejima, 1969). The unidimensionality assumption for each depression scale was assessed with a single-factor model in Mplus 6.11 (Muthén and Muthén, 1998–2007) using conventional criteria for acceptable model fit: confirmatory fit index (CFI) > 0.95, Tucker Lewis Index (TLI) > 0.95, and root mean squared error of approximation (RMSEA) < 0.08 (Reeve *et al*, 2007). When the assumption of unidimensionality was questionable, we accounted for any residual correlations among symptom items in Multiple Indicators Multiple Causes (MIMIC) modeling. Moreover, all DIF results were also confirmed in Mplus with MIMIC models (Jones, 2006).

The estimation of depression, θ, was then used as an input for the binary and ordinal logistic regression analyses. Three models were formed. The first was for the probability of endorsing the item in relation to the level of θ (model 1). The second included a term for group membership (model 2), and the third added the interaction between θ and group membership (model 3; Juhel and Gaillot, 2012). DIF was present if model 3 was significantly better than model 1. Results are presented here if the likelihood ratio (LR) chi square p-value was less than 0.01, and the McFadden R^2 was greater than 0.02. Two types of DIF have been established, non-uniform and uniform (Crane *et al*, 2006). In uniform DIF, the influence or interference is the same across all levels of depressive symptoms. In non-uniform DIF, the level of interference changes across the latent continuum of depressive symptom severity (Gibbons *et al*, 2009). A significant difference between models 1 and 2 indicated uniform DIF, and a significant difference between models 2 and 3, non-uniform DIF.

Finally, DIF presents with differing degrees of severity; because statistical power is dependent on sample size, a trivial but non-zero difference in population parameters will be found to be statistically significant given a large enough sample. We evaluated the impact of DIF on depressive symptoms measures, by comparing the original IRT scores on each depressive symptoms measure, to the constructed score that accounted for all sources of DIF. Using IRT, we estimated both the depression level and the standard error of measurement (SEM) for each individual. For each measure in each study, we subtracted the original IRT depression score from the final score that accounted for all sources of DIF and looked for changes larger than the median SEM of the original score. We have referred to such changes as salient DIF (Crane *et al*, 2010; Gibbons *et al*, 2009).

Results

Alzheimer's Disease Research Center at Duke University

The GDS sum scores ranged from 0 to 12, with a mean of 1.4 (*SD* = 2.0; 9% scored >4). The single-factor model had excellent fit, with a CFI of 0.97, TLI of 0.97, and RMSEA of 0.03. The items "Do you feel full of energy," "Do you have more problems with memory than most," and "Do you feel happy most of the time" presented with DIF because of the CDR (all LR χ^2: $p < 0.01$, $R^2 = 0.06$, 0.06, and 0.04, respectively). "Do you feel full of energy" also had DIF because of the GCP (LR χ^2: $p < 0.01$, $R^2 = 0.07$). The impact of DIF on the GDS IRT scores was not *salient*, with the maximum changes of 0.14 for the CDR and or 0.28 for the GCP, compared with a median SEM of 0.66 for the measure. The impact is illustrated in Figure 1, where all changes are well within the vertical bars indicating the SEM. MIMIC modeling also failed to uncover salient DIF (results not shown for MIMIC models).

University of California, Davis' Alzheimer's Disease Center Longitudinal Cohort

Here, the mean GDS score was 2.4 (*SD* 2.9; range 0–15; 18% scored >4). The single-factor model had excellent fit, with a CFI, TLI, and RMSEA values of 0.98, 0.97, and 0.04, respectively. The item "Do you have more problems with memory than most" presented with uniform DIF, and this was true for all three cognitive groupings: (i) consensus diagnosis of no cognitive impairment versus questionable to demented (LR χ^2: $p < 0.01$, $R^2 = 0.03$); (ii) CDR rating of ≤0.5 versus 1–3 (LR χ^2: $p < 0.01$, $R^2 = 0.05$); and (iii) GCP (LR χ^2: $p < 0.01$, $R^2 = 0.02$). There was one other item that presented with DIF, only because of the CDR, “Have you dropped many of your activities and interests” (LR χ^2: $p < 0.01$, $R^2 = 0.03$). The impact of each of these factors on the IRT GDS score was not *salient*, as the median SEM was 0.56, but the maximum differences between the scores that accounted for DIF and the original score were 0.21 for consensus diagnosis, 0.35 for the CDR, and 0.12 for the GCP (Figure 1). Validation models using the MIMIC model approach also resulted in no *salient* DIF.

In Figure 2, the top panel presents the probability of endorsing the item “Do you have more problems with memory than most” for the GCP groups. Those subjects with GCP scores above the national average have a lower probability of endorsing memory problems, except at the highest levels of depressive symptoms. This might be interpreted as the impact of depression on cognition, which has an equalizing effect on the probability of endorsing memory problems, but only depressive symptoms become severe. The bottom panel shows the expected total GDS score at each level of depression for each GCP group. The two curves are virtually identical, indicating that in a 15-item scale, one item with DIF, with a magnitude of $R^2 = 0.02$, does not have a meaningful impact at the overall test level.

Washington Heights/Hamilton Heights-Inwood Columbia Aging Project

The mean score for the 10-item CES-D was 2.0 (SD 2.1) and ranged from 0 to 10. The IRT model required a residual correlation between the two positively worded items (“Happy” and “Enjoyed life”) to achieve acceptable CFI, TLI, and RMSEA fit indices of 0.96, 0.95, and 0.06, respectively. When compared with the CES-D scores calculated with the assumption of unidimensionality, this model showed reduced factor loadings for the two correlated items. The correlation between the two IRT scores was 0.99, so we conducted analyses with the assumption of unidimensionality but also ran confirmatory analyses using MIMIC modeling. When grouping subjects by CDR 0 versus 0.5 and higher, no items were flagged for DIF. Grouping subjects by CDR <1 versus 1 and higher and also by GCP each resulted in two different pairs of items being flagged for DIF based on the LR criterion. However, none of these items reached a significant magnitude, so we identified no significant DIF either because of CDR categorization or GCP. The MIMIC model accounting for a residual correlation between the two positively worded items also indicated no significant DIF for either grouping variable.

Neurocognitive Outcomes of Depression in the Elderly

As previously noted, the *NCODE* examined depression with both the 20-item CES-D (mean 22.7, SD 16.4, range 0–58) and the MADRS (mean 23.9, SD 9.2, range 0–54). The single-factor model for the CES-D presented with acceptable fit (CFI: 0.98, TLI: 0.98, RMSEA: 0.06). Two items were flagged with DIF because of GCP, “I did not feel like eating; my appetite was poor” (LR χ^2: $p < 0.01$, $R^2 = 0.03$) and lonely (LR χ^2: $p < 0.01$, $R^2 = 0.03$). However, the impact on the total score was not *salient*; the largest change was 0.15, and the median SEM was 0.24 (Figure 1). The MIMIC model also produced no salient DIF.

In the MADRS, the single-factor model did not have acceptable fit (CFI: 0.93, TLI: 0.98, RMSEA: 0.10). Residual correlations were added between “Pes-simistic thoughts” and “Suicidal thoughts,” and between “Reported sadness” and “Reduced appetite,” to form a model with acceptable fit (CFI: 0.97, TLI: 0.99, RMSEA: 0.07). When compared with the unidimensional model, the standardized factor loadings in this model changed by, at most, 0.04 units. The two scores were correlated 0.99. We proceeded with the assumption of unidimensionality but also ran confirmatory analyses. As with the CES-D, the item concerning appetite ("Reduced appetite") reflected item bias because of GCP (LR χ^2: $p <$ 0.01, $R^2 = 0.02$), and again, the impact was not *salient*, with the largest change 1.18 and the median SEM 0.37 (Figure 1). In the confirmatory MIMIC models, there was no statistically significant DIF.

Discussion

We examined the impact of cognitive function on several common instruments used to assess depressive symptoms: the GDS, two versions of the CES-D, and the MADRS. Overall, we found that if people can understand the items enough to be administered the measure(s), researchers and clinicians can be confident that the depression scores obtained from these scales are not biased when administered to older adults who differ by levels of cognitive function. While statistically significant, DIF was found for some items on most of the measures, the overall impact on depressive symptoms scores for each of the scales was negligible.

The best-performing scale appeared to be the 10-item CES-D, which presented with little or no evidence of item bias, even when contrasting groups of older adults with normal cognitive ability and those with dementia. Other scales had at least one item with DIF. As might be expected, when grouping older adults by cognitive ability, items identified with DIF included depressive symptoms associated with memory. The most robust evidence emerged for the GDS item, "Do you feel you have more problems with memory than most." This item presented with DIF for scores dichotomized into categories of above and below the mean GCP (*UC Davis ADC*), for the CDR grouping of "no dementia" versus "mild to severe dementia" (Duke ADRC and *UC Davis ADC*), and for consensus diagnosis of no cognitive impairment versus questionable to demented (*UC Davis ADC*). It is worth noting here that, despite the observation by Tang *et al* (2005) of "no DIF" when grouping the GDS by cognitive impairment, these authors did report significant misfit for the GDS memory item. Using Rasch analysis, the OUTFIT statistic was >1.3, suggesting that the item misfit was associated with unidimensionality. Yet, for the current study, even though the GDS item met our criteria for significant DIF in four of the five instances, its effect on the total IRT score was minor. There were two other cognitive function questions that exhibited no bias: "Concentration difficulties" from the MADRS and "Keeping my mind on what I was doing" from the 20-item CES-D (*NCODE*). These items are more related to cognitive abilities associated with executive function, such as concentration and attention. Taken together, it may be that depression items related to memory work differently for individuals with varying levels of cognitive impairment, but items related to concentration and attention are less affected by cognition. However, this would need to be verified by administering all three depression symptom measures to a group of older adults who exhibit a broader range of depressive symptoms.

Our primary analytic strategy in this study was the hybrid logistic regression/IRT approach. In this iterative two-step procedure, the latent trait is estimated using structural equations modeling, and DIF is assessed using logistic regression. In the next step, the latent trait is re-estimated using separate parameters for items with DIF, and then, DIF is re-assessed. This process continues as an iterative process until a stable set of items present with DIF. We usually use it as our primary method because it is easy to assess non-uniform DIF for covariates that are not dichotomous and to account for non-uniform DIF for more than one covariate at a time. An alternative approach that is commonly used to assess DIF is the MIMIC structural equations model (Jreskog and Goldberger, 1975), in which the covariates affect the response via a latent variable only (Skrondal and Rabe-Hesketh, 2005). Of particular relevance to two of the depression measures in this study is that MIMIC models can include residual correlations between items. We confirmed all our analyses using MIMIC modeling. These validation results were consistent with results from the hybrid logistic regression/IRT approach, increasing our confidence in our conclusion that the impact of DIF was minimal.

It should be noted that a finding of no *salient* DIF is not the same as establishing measurement invariance. There are several types of measurement invariance that can be established, but in each, item parameters must be the same in the two groups. In DIF testing, we allow the parameters for some of the item to differ between the two groups. For the finding of no *salient* DIF, we have computed scores partly on the basis of differing parameters and found that those scores do not change much.

Some limitations should be mentioned. The frequency of cognitive impairment in these studies was fairly low. However, the studies included in the analysis recruited participants with a broad range of cognitive ability. Our findings are relevant to the typical range of cognitive ability observed in other studies and (Gross *et al*, 2014a) in the general community-dwelling population. Thus, our results should be informative to a majority of epidemiologic studies of cognitive aging. The process of selection for participation in each of these studies may have led to samples for whom cognitive status does not affect responses on depressive symptom measures. Even once enrolled, a depressive symptom scale may have been less likely to have been administered if the person was showing signs of cognitive impairment. Except for *NCODE*, the levels of depressive symptoms were also fairly low. Our findings may not apply to more severely impaired populations, and the GDS and 10-item CES-D findings may differ in a more depressed population. Finally, lack of DIF impact should be replicated in a population with more heterogeneity with regard to frequency of cognition impairment and depression.

Conclusion

When investigating DIF in multiple samples across commonly administered depressive symptoms scales, we found few instances of DIF related to cognitive status. Depressive symptoms items that had significant for DIF based on cognitive status included items associated with memory complaints, appetite loss, lack of energy, mood, and anhedonia. However, the impact of DIF due to cognitive status on the total scale scores was negligible. Regardless of cognitive status, the scales measured overall depressive symptoms in the same way. Our finding may reduce concerns about using these depression measures in cognitive aging research, as relationships between depression and cognitive decline are unlikely to have been due to item bias.

Acknowledgements

This work was supported by the National Institutes of Health grants 5 T32 AG00029-35 (PI: Harvey J. Cohen, MD, Duke University Medical Center), T32 AG000261 (PI: Yaakov Stern, PhD, Columbia University Medical Center), R01 MH054846 (PI: David C. Steffens, MD, MHS, University of Connecticut), R01 AG037212 (PI: Richard Mayeux, MD, MSc, Columbia University Medical Center), R01 AG029672 (PI: Paul K. Crane, MD, MPH, University of Washington), R13 AG030995 (PI: Dan M. Mungas, PhD, University of California, Davis), K23MH087741 (PI:Potter), P30 AG028377 (PI: Kathleen A. Welsh-Bohmer, PhD, Duke University Medical Center), P30 AG10129 (PI: Charles S. DeCarli, MD, University of California, Davis), P50 AG05136 (PI: Thomas Montine, MD, University of Washington), U01AR057929-S1 (PI: Alan Jette, Boston University), and Kessler Foundation.

References

Andresen EM, Malmgren JA, Carter WB, Patrick DL. Screening for depression in well older adults: evaluation of a short form of the CES-D (Center for Epidemiologic Studies Depression Scale). Am J Prev Med. 1994; 10:77–84. [PubMed: 8037935]

Carvalho JO, Tommet D, Crane PK, et al. Deconstructing racial differences: the effects of quality of education and cerebrovascular risk factors. J Gerontol B Psychol Sci Soc Sci. 2015; 70:545–556. [PubMed: 25098527]

Choi SW, Gibbons LE, Crane PK. lordif: an R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. J Stat Soft. 2011; 39:1–30.

Crane PK, Gibbons LE, Jolley L, et al. Differential item functioning related to education and age in the Italian version of the Mini-mental State Examination. Int Psychogeriatr. 2006; 18:505–515. [PubMed: 16478571]

Crane PK, Gibbons LE, Willig JH, et al. Measuring depression and depressive symptoms in HIV-infected patients as part of routine clinical care using the 9-item patient health questionnaire (PHQ-9). AIDS Care. 2010; 22:874–885. [PubMed: 20635252]

Dmitrieva NO, Fyffe D, Mukherjee S, et al. Demographic characteristics do not decrease the utility of depressive symptoms assessments: examining the practical impact of item bias in four heterogeneous samples of older adults. Int J Geriatr Psychiatry. 2014; 30:88–96. [PubMed: 24737612]

Engedal K, Kvaal K, Korsnes M, et al. The validity of the Montgomery–Asberg depression rating scale as a screening tool for depression in later life. J Affect Disord. 2012; 141:227–232. [PubMed: 22464007]

Gibbons LE, McCurry S, Rhoads K, et al. Japanese–English language equivalence of the cognitive abilities screening instrument among Japanese-Americans. Int Psychogeriatr. 2009; 21:129–137. [PubMed: 18947456]

Gilley DW, Wilson RS. Criterion-related validity of the Geriatric Depression Scale in Alzheimer's disease. J Clin Exp Neuropsychol. 1997; 19:489–499. [PubMed: 9342685]

Gross AL, Jones RN, Fong TG, Tommet D, Inouye SK. Calibration and validation of an innovative approach for estimating general cognitive performance. Neuroepidemiology. 2014a; 42:144–153. [PubMed: 24481241]

Gross A, Sherva R, Mukherjee S, et al. Calibrating longitudinal cognitive performance in alzheimer's disease across diverse neuropsychological batteries and datasets. Alzheimers and Dementia. 2014b; 10:746.

Grunebaum MF, Oquendo MA, Manly JJ. Depressive symptoms and antidepressant use in a random community sample of ethnically diverse, urban elder persons. J Affect Disord. 2008; 105:273–277. [PubMed: 17532052]

Hertzog C, van Alastine J, Usala PD, Hultsch DF, Dixon R. Measurement properties of the Center for Epidemiological Studies Depression Scale (CES-D) in older populations. Psychological Assessment: J Consult Clin Psych. 1990; 2:64–72.

Hinton L, Carter K, Reed BR, et al. Recruitment of a community-based cohort for research on diversity and risk of dementia. Alzheimer Dis Assoc Disord. 2010; 24:234–241. [PubMed: 20625273]

Hulin CL. A psychometric theory of evaluations of item and scale translations: fidelity across languages. J Cross Cult Psychol. 1987; 18:115–142.

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. Med Care. 2006; 44:S124–S133. [PubMed: 17060819]

Jorm AF. History of depression as a risk factor for dementia: an updated review. Aust N Z J Psychiatry. 2001; 35:776–781. [PubMed: 11990888]

Juhel J, Gaillot AC. Structural validity and age-based differential item functioning of the French Nottingham Health Profile in a sample of surgery patients. Adv Psychol Stud. 2012; 1:14–21.

Jreskog K, Goldberger A. Estimation of a model of multiple indicators and multiple causes of a single latent variable. J Am Stat Assoc. 1975; 10:631–639.

Manly JJ, Bell-Mcginty S, Tang M-X, et al. Implementing diagnostic criteria and estimating frequency of mild cognitive impairment in an urban community. Arch Neurol. 2005; 62:1739–1746. [PubMed: 16286549]

McGivney SA, Mulvihill MM, Taylor B. Validating the GDS depression screen in the nursing home. J Am Geriatr Soc. 1994; 42:490–492. [PubMed: 8176142]

Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry. 1979; 134:382–389. [PubMed: 444788]

Montorio I, Izal M. The geriatric depression scale: a review of its development and utility. Int Psychogeriatr. 1996; 8:103–112. [PubMed: 8805091]

Morris JC. The clinical dementia rating (CDR): current version and scoring rules. Neurology. 1993; 43:2412–2414. [PubMed: 8232972]

Mottram P, Wilson K, Copeland J. Validation of the Hamilton Depression Rating Scale and Montgomery and Asberg Rating Scales in terms of AGECAT depression cases. Int J Geriatr Psychiatry. 2000; 15:1113–1119. [PubMed: 11180467]

Mungas D, Beckett L, Harvey D, et al. Heterogeneity of cognitive trajectories in diverse older persons. Psychol Aging. 2010; 25:606–619. [PubMed: 20677882]

Muthén, LK.; Muthén, BO. 1998-2007. Mplus: Statistical Analysis with Latent Variables. 5.1. Los Angeles, CA; Muthén & Muthén:

Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. Appl Psychol Meas. 1977; 1:385–401.

Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care. 2007; 45:S22–S31. [PubMed: 17443115]

Samejima, F. Estimation of Latent Ability Using a Response Pattern of Graded Scores (Psychometric Monograph No. 17). Richmond, VA; Psychometric Society: 1969.

Scott NW, Fayers PM, Aaronson NK, et al. Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. Health Qual Life Outcomes. 2010; 8:1–9. [PubMed: 20053296]

Sheikh JI, Yesavage JA. Geriatric Depression Scale (GDS): recent evidence and development of a shorter version. Clinical Gerontologist: J Ment Health Aging. 1986; 5:165–172.

Skrondal A, Rabe-Hesketh S. Structural equation modeling: categorical variables. Entry for the Encyclopedia of Statistics in Behavioral Science. 2005 doi: 10.1002/0470013192.bsa596.

Steffens DC, Welsh-Bohmer KA, Burke JR, et al. Methodology and preliminary results from neurocognitive outcomes of depression in the elderly study. J. Geriatr Psychiatry Neurol. 2004; 17:202–211. [PubMed: 15533991]

Tang WK, Wong E, Chiu HF, Lum CM, Ungvari GS. The Geriatric Depression Scale should be shortened: results of Rasch analysis. Int J Geriatr Psychiatry. 2005; 20:783–789. [PubMed: 16035120]

Tukey, JW. Exploratory Data Analysis. Reading, MA; Addison-Wesley Publishing Co: 1977.

Williams, JW.; Plassman, BL.; Burke, J., et al. Evidence Report/Technology Assessment No. Rockville, MD; Agency for Healthcare Research and Quality: 2010. 193. Preventing Alzheimer's Disease and Cognitive Decline.

Key points

- Our objective was to examine the significance and practical impact of psychometric bias due to cognitive status, across four large heterogeneous samples of older adults.
- The three depressive symptoms scales in these four studies measured depression in the same way, regardless of cognitive status.
- These findings may reduce concerns about the utilization of these depression measures in cognitive aging research, as relationships between depression and cognitive decline are unlikely to have been due to item bias, at least in the ways that were measured in the current study.

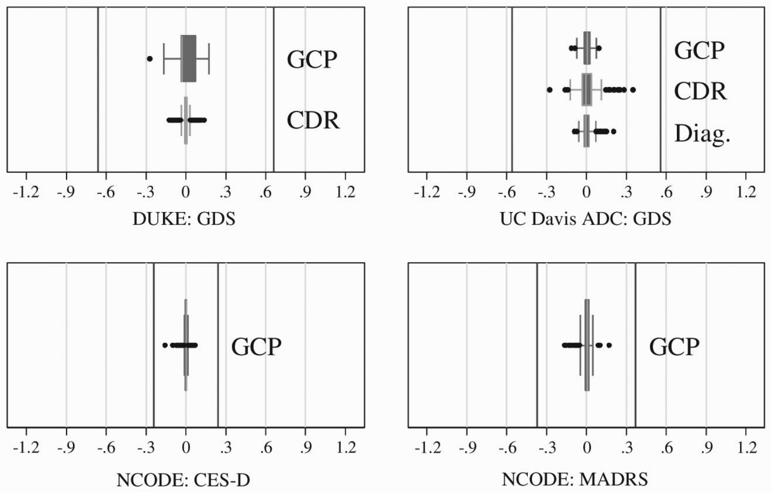

Figure 1.

Box plots of the changes in the item response theory based depression scores after accounting for differential item functioning (DIF). The plots show the difference between unadjusted scores and scores accounting for DIF due to cognitive category. If DIF had no impact for an individual, that observation should lie at zero. The grayed boxes represent the interquartile range, and the whiskers signify the upper and lower adjacent values as defined by Tukey (Tukey, 1977). Observations more extreme than the upper and lower adjacent values are outliers, which are represented by dots. Vertical lines are placed at one standard error of measurement for each scale in each sample, and observations outside the lines would indicate the presence of *salient* DIF. GCP = General Cognitive Performance. CDR = Clinical Dementia Rating. Diag. = Diagnostic category. There was no statistically significant DIF detected in the *WHICAP* CES-D 10.

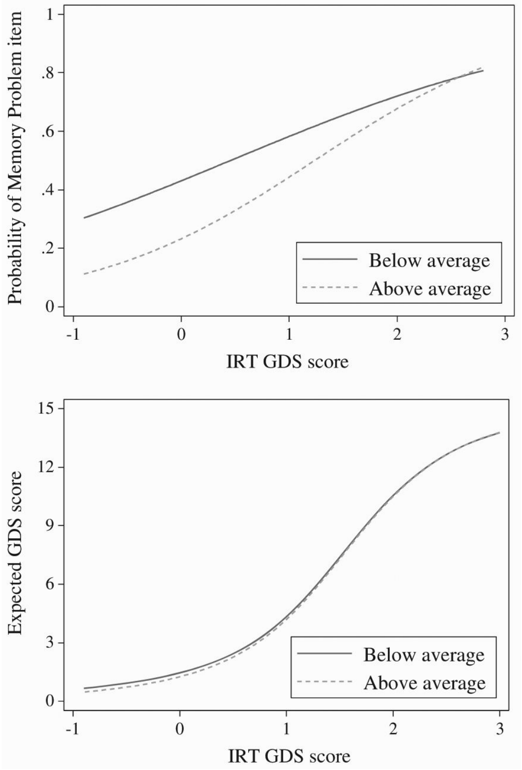

Figure 2.

The probability of endorsing “Do you have more problems with memory than most” (top panel) and the expected total Geriatric Depression Scale score (bottom panel), at each level of depression (as estimated by the item response theory based Geriatric Depression Scale score), for University of California, Davis’ Alzheimer’s Disease Center Longitudinal Cohort participants with General Cognitive Performance below and above the US average.

Table 1

Participant characteristics

	Duke ADRC (*N*=511)	**UC Davis ADC (*N* = 620)**	**WHICAP (*N* = 2137)**	**NCODE (*N* = 290[a])**
Characteristic	***M* (SD) or %**	***M* (SD) or %**	***M* (SD) or %**	***M* (SD) or %**
Age (years)	72.1 (9.0)	75.8 (7.5)	77.0 (7.1)	69.2 (6.3)
Female	60.9	58.6	67.1	64.8
Race/ethnicity				
Black	22.9	23.9	33.1	11.0
Caucasian	77.1	48.7	30.1	89.0
Hispanic	—	23.2	36.8	—
Other	—	4.2	—	—
Education (years)	15.5 (2.8)	13.1 (4.3)	10.2 (4.9)	14.7 (2.5)
Tested in English (versus Spanish)	100.0	88.3	66.3	100.0
Clinical Dementia Rating Scale				
0	61.1	40.4	70.4	—
0.5	32.1	44.2	19.6	—
1–3	6.8	15.4	10.0	—
General cognitive performance score above US average[b]	72.2	47.1	44.7	85.2
Diagnosis of no cognitive impairment (versus questionable to dementia)	—	33.6	—	—

DUKE ADRC, Alzheimer's Disease Research Center at Duke University; MADRS, Montgomery–Åsberg Depression Rating Scale; NCODE, Neurocognitive Outcomes of Depression in the Elderly Study; UC DAVIS ADC,University of California, Davis' Alzheimer's Disease Center Longitudinal Cohort.

[a] MADRS data were available for *n* = 153.

[b] Computed from each study's neuropsychological battery and linked to results from a survey of the US population, using calibration methods found in Gross *et al*. (2014a).