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

Natalia O. Dmitrieva¹, Denise Fyffe², Shubhabrata Mukherjee³, Robert Fieò⁴, Laura B. Zahodne⁴, Jamie Hamilton⁴, Guy G. Potter¹,5, Jennifer J. Manly⁴, Heather R. Romero⁵,6, Dan Mungas⁷ and Laura E. Gibbons³

¹Center for the Study of Aging and Human Development, Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Durham, NC, USA
²Kessler Foundation Research Center, Department of Physical Medicine and Rehabilitation, New Jersey Medical School, Rutgers, the State University of New Jersey, West Orange, NJ, USA
³General Internal Medicine, University of Washington, Seattle, WA, USA
⁴Cognitive Neuroscience Division, Department of Neurology and Taub Institute for Research on Alzheimer’s Disease and the Aging Brain, Columbia University, New York, NY, USA
⁵Joseph and Kathleen Bryan Alzheimer’s Disease Research Center, Duke University Medical Center, Durham, NC, USA
⁶Department of Psychology, University of Notre Dame, Notre Dame, IN, USA
⁷Department of Neurology, University of California, Davis, University of California, Davis Medical Center, Sacramento, CA, USA

Correspondence to: N. O. Dmitrieva, PhD, E-mail: natalia.dmitrieva@duke.edu

Objective: Previous studies have identified differential item function (DIF) in depressive symptoms measures, but the impact of DIF has been rarely reported. Given the critical importance of depressive symptoms assessment among older adults, we examined whether DIF due to demographic characteristics resulted in salient score changes in commonly used measures.

Methods: Four longitudinal studies of cognitive aging provided a sample size of 3754 older adults and included individuals both with and without a clinical diagnosis of major depression. Each study administered at least one of the following measures: the Center for Epidemiologic Studies Depression scale (20-item ordinal response or 10-item dichotomous response versions), the Geriatric Depression Scale, and the Montgomery–Åsberg Depression Rating Scale. Hybrid logistic regression-item response theory methods were used to examine the presence and impact of DIF due to age, sex, race/ethnicity, and years of education on the depressive symptoms items.

Results: Although statistically significant DIF due to demographic factors was present on several items, its cumulative impact on depressive symptoms scores was practically negligible.

Conclusions: The findings support substantive meaningfulness of previously reported demographic differences in depressive symptoms among older adults, showing that these individual differences were unlikely to have resulted from item bias attributable to demographic characteristics we examined.

Introduction

Diagnostic issues are of critical importance in geriatric depression, a condition that is underdiagnosed and undertreated (Alexopoulos, 2005). Whereas effective treatment of depression among older adults has been linked to improved quality of life and reduced cost of health care (Alexopoulos et al., 2001), untreated depression has been associated with poorer cognitive and physical functioning, and increased suicide rate (Blazer,
characteristics, including age (e.g., Kessler research, depressive symptoms vary by demographic 2003). As has been shown by an extensive body of depression underdiagnosis. (DIF; Camilli and Shepard, 1994). In the case of depression, the latter source is known as differential item functioning (DIF; Camilli and Shepard, 1994). The former source of differences suggests true demographic dissimilarity, whereas the chosen reference group. The former source of differences suggests true demographic dissimilarity, whereas the latter source is known as differential item functioning (DIF; Camilli and Shepard, 1994). In the case of depressive symptoms, such problems in assessment may lead to false positive diagnoses, thereby increasing public cost of care, or false negative diagnoses, thereby exacerbating depression underdiagnosis.

A number of studies have reported statistically significant differential DIF in commonly used depressive symptoms measures. Several studies investigating sex-based DIF among older adults on the Geriatric Depression Scale (GDS) and the Center for Epidemiologic Studies Depression scale (CES-D) have demonstrated that compared with men with the same level of depressive symptoms, women are less likely to endorse the GDS item “Hopeless” (e.g., Broekman et al., 2008), but more likely to endorse the CES-D item “Crying” (e.g., Cole et al., 2000). Others have shown that compared with Whites with the same level of depressive symptomatology, Blacks or African Americans are more likely to endorse CES-D items “People were unfriendly” and “People disliked me,” (e.g., Yang et al., 2009), and Mexican Americans are more likely to endorse some of the positively worded items, including “Happy” (e.g., Kim et al., 2009).

Aside from these relatively consistent findings, much of the evidence for demographic-based DIF varies, such that items may reach the threshold for statistical significance in some studies, but not in others. This is the case with age-based DIF on GDS items “Dropped activities,” “Stay home,” “Memory problems,” “Worthless,” “Energy,” and CES-D items “Failure,” “Hopeful,” “Bothered” (Cole et al., 2000; Grayson et al., 2000; Broekman et al., 2008; Marc et al., 2008; Yang et al., 2009); sex-based DIF on GDS items “Good spirits,” “Memory problems,” “Afraid,” and CES-D items “People were unfriendly,” “People disliked me,” “Failure,” “Talked less,” “Effort,” “Fearful,” “Restless sleep” (Stommel et al., 1993; Grayson et al., 2000; Broekman et al., 2008; Marc et al., 2008; Yang et al., 2009); race-based DIF on the GDS item “Memory problems” (Marc et al., 2008); and race/ethnicity-based DIF on the CES-D items “As good as other people,” “Hopeful,” and “Enjoyed life” (Kim et al., 2009).

The first objective of current manuscript was to assess item-level DIF in commonly used depressive symptoms measures among large and diverse samples of older adults. Specifically, we utilized data from four longitudinal studies of cognitive aging to examine test bias attributable to age, sex, race/ethnicity, and years of education in the following depressive symptoms instruments: GDS, CES-D (20-item ordinal response and 10-item dichotomous response versions), and the Montgomery–Åsberg Depression Rating Scale (MADRS)). Following item-level DIF detection, an important, but relatively infrequent step is to examine the degree to which statistically significant DIF results are salient or meaningful (Crane et al., 2010). Thus, our second objective was to compare DIF-accounted depressive symptoms scores with the original scores, and to identify instances when the difference in the two scores was large enough to be practically meaningful.

Methods

Participants

Differential item function in depressive symptoms was assessed among participants (N = 3754) in four studies of cognitive aging as follows: (1) the Joseph and Kathleen Bryan Alzheimer’s Disease Research Center at Duke University (Duke ADRC; n = 511); (2) the Spanish and English Neuropsychological Assessment Scales at the University of California, Davis’ Alzheimer’s Disease Center (UCD ADC; n = 620); (3) the Washington Heights/Hamilton Heights-Inwood Columbia Aging Project (WHICAP; n = 2045); and (4) the Neurocognitive Outcomes of Depression in the Elderly Study (NCOED; n = 578). We examined data from the first assessment of depressive symptoms in each study. Each study’s methodology is described in previously published work (Duke ADRC: Romero et al., 2010; Carvalho et al., submitted, UCD ADC: N. O. Dmitrieva et al.
Differential item function in depressive symptoms

Hinton et al., 2010, WHICAP: Manly et al., 2005, NCODE: Steffens et al., 2004). For a brief overview, refer to Supporting Information. All study procedures were approved by their respective Institutional Review Boards.

Measures

GDS. Duke ADRC and UCD ADC utilized the 15-item GDS (Sheikh and Yesavage, 1986), a validated and reliable scale with dichotomous response categories that was developed to assess depressive symptoms in the older adult population (Lyness et al., 1997). The average GDS score was 1.4 (standard deviation (SD) = 2.0, range 0–12) in Duke ADRC and 2.4 (SD = 2.9, range 0–15) in UCD ADC.

CES-D. The CES-D (Radloff, 1977) is a commonly used measure of depressive symptoms among adults of all ages; both long and short versions of the CES-D have been validated for use among older adults (e.g., Lyness et al., 1997). A 10-item dichotomous response version was used in the WHICAP (Mean (M) = 1.9, SD = 2.1, range 0–10), whereas the 20-item ordinal response version was used in NCODE (M = 22.7, SD = 16.4, range 0–58).

MADRS. The NCODE also administered the MADRS (Montgomery and Åsberg, 1979) to a subsample of participants (n = 389). The MADRS is a clinician-rated 10-item ordinal-response scale validated among older adults (Mottram et al., 2001). The average MADRS score was 23.9 (SD = 16.4, range 0–54).

Demographic variables. Participant characteristics are detailed in Table 1. For consistency across studies, we used race and ethnicity data collected by each study to classify participants according to the Office of Management and Budget Standards for maintaining, collecting, and presenting data on race and ethnicity (Office of Management and Budget, 1997), which are also used by the United States Census Bureau.

Analytic strategy

The general analytic strategy was the following: (1) to use item-response theory (IRT) to assess DIF sequentially due to age, sex, race/ethnicity, and years of education; and (2) to identify DIF salience. Where applicable, we also assessed DIF due to test language. As IRT requires unidimensionality, we preceded DIF analyses by first confirming the unidimensionality assumption in depressive symptoms scales.

Unidimensionality. The unidimensionality assumption was assessed with a single factor model in Mplus 6.11 (Muthén & Muthén, Los Angeles, CA, USA) (Muthén and Muthén, 1998–2010) using conventional criteria for acceptable model fit: Comparative Fit Index (CFI) > 0.95, Tucker-Lewis index (TLI) > 0.95, and Root Mean Square Error of Approximation (RMSEA) < 0.08 (Reeve et al., 2007). When the assumption of unidimensionality was questionable, results were confirmed in Mplus with Multiple Indicators Multiple Table 1  Participant characteristics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Duke ADRC (N=511)</th>
<th>UCD ADC (N=620)</th>
<th>WHICAP (N=2045)</th>
<th>NCODE (N=578)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M (SD) or %</td>
<td>M (SD) or %</td>
<td>M (SD) or %</td>
<td>M (SD) or %</td>
</tr>
<tr>
<td>Sex (1 = Female)</td>
<td>60.9 (9.0)</td>
<td>58.6 (7.5)</td>
<td>67.4 (7.1)</td>
<td>68.2 (7.0)</td>
</tr>
<tr>
<td>Age (years)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Race/ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black or African American</td>
<td>22.9</td>
<td>23.9</td>
<td>34.3</td>
<td>11.6</td>
</tr>
<tr>
<td>White</td>
<td>77.1</td>
<td>48.7</td>
<td>31.2</td>
<td>88.4</td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td></td>
<td>23.2</td>
<td>34.4</td>
<td>—</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>4.2</td>
<td></td>
<td>—</td>
</tr>
<tr>
<td>Education (years)</td>
<td>15.5 (2.8)</td>
<td>13.1 (4.3)</td>
<td>10.2 (4.9)</td>
<td>14.0 (2.9)</td>
</tr>
<tr>
<td>Tested in English (vs. Spanish)</td>
<td>100.0</td>
<td>88.3</td>
<td>65.3</td>
<td>100.0</td>
</tr>
<tr>
<td>Clinical Dementia Rating Scale†</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>61.1</td>
<td>40.4</td>
<td>70.1</td>
<td>—</td>
</tr>
<tr>
<td>0.5</td>
<td>32.1</td>
<td>44.2</td>
<td>19.7</td>
<td>—</td>
</tr>
<tr>
<td>1–3</td>
<td>6.8</td>
<td>15.4</td>
<td>10.2</td>
<td>—</td>
</tr>
</tbody>
</table>

Duke ADRC, Joseph and Kathleen Bryan Alzheimer’s Disease Research Center at Duke University; UCD ADC, Spanish and English Neuropsychological Assessment Scales at the University of California, Davis’ Alzheimer’s Disease Center; WHICAP, Washington Heights/Hamilton Heights-Inwood Columbia Aging Project; NCODE, Neurocognitive Outcomes of Depression in the Elderly Study; MADRS, Montgomery–Åsberg Depression Rating Scale; M, mean; SD, standard deviation.

†Includes only Whites, Blacks, or African Americans tested in English and Hispanics tested in Spanish.

‡MADRS data were available for 389 participants.

§NCODE did not administer the Clinical Dementia Rating Scale.
Causes (MIMIC) models, which accounted for residual correlations among symptom items (Jones, 2006).

**IRT-based depression scores.** Item-response theory based depression scores were computed with PARSCALE 4.1 (Scientific Software International, Chicago, IL, USA) (Muraki and Bock, 2003), using Samejima’s graded response model (Samejima, 1969) for ordinal responses.

**DIF detection.** We assessed item-level uniform (differences present across all levels of depressive symptoms) and nonuniform (differences that vary by depressive symptoms level) DIF using the hybrid ordinal logistic regression-IRT method for DIF detection, with the Stata (StataCorp, 2011, StataCorp LP, College Station, TX, USA) command -difwithpar- (Crane et al., 2006, Gibbons et al., 2009). DIF-free items were used as anchors, and group-specific item parameters were estimated for items with DIF. DIF significance was initially set to a liberal $p \leq 0.05$ criterion. Due to the large sample and the small number of dichotomous items, we found a $p \leq 0.01$ threshold was needed in WHICAP to retain enough anchor items.

In Duke ADRC and NCODE samples, DIF significance was assessed sequentially, starting with age, followed by sex, race, and education. Age was dichotomized at each study-specific median ($\leq$72 years old ($M = 65$) versus $> 72$ years old ($M = 80$) in Duke ADRC; and $\leq$69 years old ($M = 64$) versus $> 69$ years old ($M = 76$) in NCODE). We then accounted for sex-based DIF, race-based DIF (Black or African American versus White), and education years-based DIF ($\leq 15$ versus $> 15$ years in Duke ADRC, and $\leq 14$ versus $> 14$ years in NCODE).

In UCD ADC and WHICAP, DIF detection started with age ($\leq 75$ years old ($M = 70$) versus $> 75$ years old ($M = 82$) in UCD ADC; $\leq 75$ years old ($M = 71$) versus $> 75$ years old ($M = 83$) in WHICAP), then sex. Sequential DIF detection due to education, race/ethnicity, and test language was not feasible, because these factors were highly correlated in UCD ADC and WHICAP. To address this, we created race/ethnicity–education–test language groups of participants specific to each study.

Four groups were created in UCD ADC: White participants assessed in English; Black or African American participants assessed in English; Hispanic or Latino participants assessed in English; and Hispanic or Latino participants assessed in Spanish. The effect of education years was examined within each race/ethnicity–test language group, by dividing each group at its median years of education: $\leq 13$ versus $> 13$ years among the White and Black or African American groups. The small sample sizes in the Hispanic or Latino assessment language groups precluded accounting for DIF due to education years within test language. Instead, we tested the effect of education years ($\leq 5$ versus $> 5$ years) among the entire group of Hispanic or Latino participants without regard to language of assessment.

Three groups were created in WHICAP: non-Hispanic White participants tested in English, non-Hispanic Black or African American participants tested in English, and Hispanic or Latino participants tested in Spanish. The effect of education years was examined within each race/ethnicity–test language group, by dividing each group at its median years of education: $\leq 5$ versus $> 5$ years within the Hispanic or Latino groups, and $\leq 13$ versus $> 13$ years among the other groups.

**Salient DIF.** Following item-level DIF detection, DIF salience was assessed by comparing the original score with the DIF-accounted depressive symptoms score, noting all instances when the difference between the two exceeded the original score’s standard error of measurement (SEM; Gibbons et al., 2009); this degree of change that has been associated with meaningful differences (e.g., Bartels et al., 2004).

**Results**

**Duke ADRC.** The single-factor model had excellent fit, with a CFI of 0.97, TLI of 0.97 and RMSEA of 0.03. The median SEM on the GDS was 0.66. There was no significant DIF due to age, although several items exhibited statistically significant DIF due to race, sex, and years of education (Supporting Information Table 1). Accounting for DIF caused slight changes in GDS group means: scores among Whites, women, and those with more than 15 years of education increased, whereas scores among Blacks, men, and those with $\leq 15$ years of education decreased.

Shifts in scores were all well within the limits of one SEM for age, sex and race (Figure 1). One participant had a salient score change due to education, a man with over 15 years of education who endorsed only the item “Better.” His estimate changed from 0.2 to $-0.8$, indicating that he was less depressed than a DIF-naïve score would indicate. Overall changes were minor, but one can observe a slight decrease in scores in the top box ($\leq 15$ years) and a slight increase in the lower box ($> 15$ years) after accounting for DIF. Though minor, this was the largest change due to DIF observed in any of the four samples (a decrease of 0.11 SD in the mean GDS for the group with fewer years of education and an increase of 0.07 SD in the group with more years...
of education). All other group mean changes in all four studies were less than 0.06 SD, and most were less than 0.02 SD. When all sources of DIF were accounted for (Figure 2), five additional people had salient DIF: four who endorsed only the item “Better” and one whose estimate also decreased by 1 SD.

**UCD ADC.** The single-factor model had excellent fit, with a CFI, TLI, and RMSEA values of 0.98, 0.97, and 0.04, respectively. The median SEM of the GDS was 0.56. There was statistically significant DIF due to age, sex, and race/ethnicity–test language group, and there was no DIF due to years of education within each race/ethnicity group (Supporting Information Table 2). No UCD ADC participant had salient DIF (Figure 2).

**WHICAP.** The single-factor model did not have acceptable fit (CFI: 0.90, TLI: 0.88, and RMSEA: 0.09). We added a residual correlation between the two positively worded items (“Happy,” and “Enjoyed life”), resulting in 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 CES-D scores from this scale correlated 0.99 with the single-factor model. Thus, we proceeded with the assumption of
unidimensionality, but also ran confirmatory analyses using MIMIC modeling. The median SEM was 0.54.

There was statistically significant DIF due to age, sex, and race/ethnicity–test language–years of education groups (Supporting Information Table 3), but the overall impact was minimal, even after accounting for all demographic sources of DIF (Figure 2). The largest change in CES-D score was 0.37. We examined DIF due to years of education separately, ignoring race/ethnicity and test language, and also saw no salient DIF. When years of education were divided at the median within each race/ethnicity group, there was no DIF among the White and Hispanic or Latino participants, and no salient DIF among Black or African American participants. Mplus validation analyses accounting for a residual correlation between the two positively worded items produced similar results (not presented).

NCODE CES-D. The single-factor model had acceptable fit (CFI: 0.98, TLI: 0.98, and RMSEA: 0.06). There was statistically significant DIF due to each demographic covariate (Supporting Information Table 4), but the impact was negligible, with all the changes well under the median SEM of 0.29 (Figure 2).

NCODE MADRS. The single-factor model did not have acceptable fit (CFI: 0.93, TLI: 0.98, and RMSEA: 0.10). Residual correlations were added between “Pessimistic 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 MIMIC models as confirmatory analyses.

There was statistically significant DIF due to demographic characteristics (Supporting Information Table 5). After accounting for all sources of DIF, we found negligible impact on group MADRS scores (Figure 2). There were 11 individuals with salient DIF. Two of the 11 participants exhibited a change of >0.5 SD on the MADRS. These participants had a MADRS score of 52 and 54, and would be considered depressed with either the original IRT score (2.6) or the DIF-adjusted score (3.2). Mplus validation analyses showed that changes in scores were similar in models with and without the residual correlation (results not presented).

Secondary analyses. Because rates of cognitive impairment vary by demographic characteristics (e.g., age), we were concerned that any significant DIF attributed to one or more of these demographic factors may be due to differences in cognitive status. We ran DIF analyses omitting participants with a score of 1 or higher on the Clinical Dementia Rating Scale (Hughes et al., 1982) in Duke ADRC, UCD ADC, and WHICAP samples (NCODE did not administer the Clinical Dementia Rating Scale). The results revealed that one participant in Duke ADRC had salient DIF due to years of education, with no participants exhibiting salient DIF due to other demographic characteristics or in other samples.

Discussion

Understanding health disparities in mental health outcomes requires the establishment of measurement equivalence (Stewart and Nápoles-Springer, 2003), which confirms that individuals are screened consistently across groups. A number of recent studies have called into question the extent to which demographic differences in depression may, in part, reflect demographic effects on psychometric properties of depressive symptoms scales. Our results demonstrate that although statistically significant bias was indeed detectable on some items, its impact on final depressive symptoms scores was practically negligible. Thus, the current study supports the practical utility of the examined scales, as well as the substantive meaningfulness of previously reported individual differences in depressive symptoms levels.

We have shown that some GDS, CES-D, and MADRS items exhibit statistically significant DIF due to age, sex, race/ethnicity, and years of education. We confirmed several previous findings (e.g., sex-based DIF on the CES-D item “Crying”, Cole et al., 2000), but not others (e.g., race-based DIF on CES-D items “People were unfriendly” and “People disliked me”; Yang et al., 2009). The presence of statistically significant item bias leads earlier researchers to recommend dropping items with DIF (e.g., Stommel et al., 1993). As explained in a more recent work (e.g., Teresi et al., 2012), dropping items with DIF is unnecessary. Readily available statistical approaches using Structural Equations Modeling or hybrid ordinal logistic regression-IRT modeling allow researchers to simultaneously account for potential sources of DIF with covariates or to output modified scale scores that account for all sources of DIF. These approaches allow for formal assessment of DIF, while retaining all available scale items for greater precision in parameter estimation.

The current study also evaluated whether significant item bias resulted in salient changes in total depressive symptoms scores. Current psychometric literature does not clearly articulate thresholds designating varying degrees of DIF significance (Yang and Jones, 2008).
Previously reported definitions of salient DIF differ by study and have included the following: item-level DIF crossing the threshold α level of 0.05 (Areán and Miranda, 1997), item-level DIF with Bonferroni-corrected p-values (Kim et al., 2009), a critical ratio of 1.25 standard error (Uebelacker et al., 2009), and a critical ratio of 2 standard error (Grayson et al., 2000). Given the importance of the overall level of depressive symptoms in research and clinical settings, we focused on the change in total score (rather than on the change in individual items), as the most practical approach to detecting salient DIF. Using the median SEM as the critical threshold, we found that statistically significant demographics-based differences in item endorsement did not yield changes in scores that were practically meaningful.

Limitations and future directions

Among the limitations of the current study, it is important to note that our comparisons of race/ethnicity-based DIF were based on data from Black or African American, Hispanic or Latino, and White participants. This means that future comparisons among older adults of different racial and ethnic backgrounds may detect salient DIF. Similarly, we only compared Spanish and English language versions, and future work should assess bias due to testing language across other commonly utilized translations of these depressive symptoms scales. We should also note that except for NCODE, the average depressive symptoms scores were relatively low, resulting in less precision for the higher levels of the depressive symptoms construct. NCODE and Duke ADRC participants completed a relatively high number of years of education (mean 14.0 and 15.5, respectively). Although UCD ADC and WHICAP participants reported a greater range of education years, education was strongly related to assessment language, precluding a test of DIF due only to years of education. Thus, it is possible that the current samples did not allow for detection of residual DIF due to fewer years of education.

Prior work demonstrates that a simple indicator of years of schooling is insufficient for understanding differences due to educational experiences among racially and ethnically diverse adults (Manly et al., 2004). A more sophisticated analysis of education-based DIF in depressive symptoms should investigate DIF due to various facets of education quality (e.g., literacy level), which more fully capture the complexity of the educational experience construct. Our findings are based on samples of older adults who exhibit a wide range of cognitive function. While secondary analyses indicate that our findings on demographics-based DIF were not due to differences in cognitive status, future work should examine cognitive status as a source of DIF in depressive symptoms items and scores.

The SEM approach is limited in that it depends on measurement precision of each depressive symptoms scale, rather than on clinically determined criteria that have been shown to signify meaningful differences in depressive symptoms. A lack of established clinically important differences on the CES-D and GDS (cf. Duru and Fantino, 2008 for MADRS) precluded us from determining whether the small demographic-based DIF found in the current study was sufficiently substantial to reach the threshold of clinically meaningful differences.

Conclusion

The current study examined item bias across four commonly used scales of depressive symptoms and across four large samples with different sampling protocols, which included participants both with and without a clinical diagnosis of depression. We demonstrated that the cumulative impacts of demographic characteristics on measurement are practically negligible, suggesting that researchers can effectively assess depressive symptoms among older adults across the four major sociodemographic characteristics.

Conflict of interest

None declared.

Key points

- In light of the growing concern with item bias in depressive symptoms measures, we determined the significance and practical impact of item bias due to demographic characteristics among four large heterogeneous samples of older adults.
- We detected statistically significant, but not practically meaningful impact of DIF due to age, sex, race/ethnicity, and years of education on four commonly used measures of depressive symptoms.
- Although statistically significant DIF due to demographic factors is detectable on several depressive symptoms items, its cumulative impact on depressive symptoms scores is practically negligible.
- The results support previously reported demographic differences in depressive symptoms, showing that these individual differences were unlikely to have resulted from item bias attributable to age, sex, race/ethnicity, or years of education.
Acknowledgements

We are grateful to Professors Carl F. Pieper and Gerda G. Fillenbaum for reviewing and providing editorial suggestions on an earlier draft of this manuscript, and to Elizabeth Sanders, for creating the figures.

This work was supported by the following National Institutes of Health grants: ST32 AG00029-35 (PI: Harvey J. Cohen, MD, Duke University Medical Center), T32 AG000261 (PI: Yaakov Stern, PhD, Columbia University Medical Center), T32 MH20004 (PI: Steven Roose, MD, 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), P30 AG028377 (PI: Kathleen A. Welsh-Bohmer, PhD, Duke University Medical Center), P30 AG10129 (PI: Charles S. DeCarli, MD, University of California, Davis), and P50 AG05136 (PI: Murray A. Raskind, MD, University of California, Washington). This work was also supported by the Kessler Foundation (to Denise Fyffe, PhD).

References


Copyright © 2014 John Wiley & Sons, Ltd.


**Supporting Information**

Additional supporting information may be found in the online version of this article at the publisher’s web site.