

## Test of item-response bias in the CES-D scale: experience from the New Haven EPESE Study

Stephen R. Cole<sup>a,b,\*</sup>, Ichiro Kawachi<sup>b</sup>, Susan J. Maller<sup>d</sup>, Lisa F. Berkman<sup>b,c</sup>

<sup>a</sup>Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02215, USA

<sup>b</sup>Department of Health and Social Behavior, Harvard School of Public Health, Boston, MA 02215, USA

<sup>c</sup>Department of Epidemiology, Harvard School of Public Health, Boston, MA 02215, USA

<sup>d</sup>Department of Educational Studies, Purdue University, West Lafayette, IN 47907, USA

Received 31 March 1999; received in revised form 28 June 1999; accepted 19 July 1999

### Abstract

We present results of item-response bias analyses of the exogenous variables age, gender, and race for all items from the Center for Epidemiologic Studies Depression (CES-D) scale using data ( $N = 2340$ ) from the New Haven component of the Established Populations for Epidemiologic Studies of the Elderly (EPESE). The proportional odds of blacks responding higher on the CES-D items “people are unfriendly” and “people dislike me” were 2.29 (95% confidence interval: 1.74, 3.02) and 2.96 (95% confidence interval: 2.15, 4.07) times that of whites matched on overall depressive symptoms, respectively. In addition, the proportional odds of women responding higher on the CES-D item “crying spells” were 2.14 (95% confidence interval: 1.60, 2.82) times that of men matched on overall depressive symptoms. Our data indicate the CES-D would have greater validity among this diverse group of older men and women after removal of the crying item and two interpersonal items. © 2000 Elsevier Science Inc. All rights reserved.

**Keywords:** Depressive symptoms; CES-D; Psychometrics; Item-response bias

### 1. Introduction

The Center for Epidemiologic Studies Depression (CES-D) scale [1] is a widely accepted 20-item self-report instrument that quantifies depressive symptoms. Valid use of a summary score for this, or any, multi-item measurement scale assumes no item-level bias by variables exogenous to the latent construct being measured [2,3]. Establishing such scale invariance is of prime importance for drawing unbiased inferences in research [4,5].

Item bias is the degree to which items that comprise a measurement scale are systematically related to various exogenous variables (e.g., age, gender, race, etc.) after conditioning on the latent variable of interest, which is often represented by the total scale score. Items may be biased in difficulty and/or discrimination. Item difficulty bias occurs when one group, for example women, responds higher on a certain item after being matched to men on the total scale score. Item discrimination bias occurs when the item difficulty bias (difference between two groups matched on the

latent variable) increases or decreases as a function of the level of the latent variable. The few reports of item bias in the biomedical literature address item difficulty [5–8].

The effect of age on the measurement properties of the CES-D remains uncertain. Previous investigators concentrated on potential factor-level biases and reported either factorial invariance [9–11] or positive age bias in at least one of the four CES-D factors (depressed affect, low positive affect, somatic complaints, and interpersonal problems) [12–14]. None of these studies intended or used methods appropriate for the assessment of item bias.

The effect of gender on the measurement properties of the CES-D is similarly uncertain. Several studies have presented both item- and factor-level data [1,9,15], suggesting a higher endorsement of depressive symptoms among women. To date only a single study has used methods appropriate to assess item bias by gender among adults administered the CES-D. Stommel et al. [16], using a series of multisample confirmatory factor analysis models on a sample of 1212 subjects, found that two CES-D items were gender biased. The “crying” item had higher endorsement among females than males, while the “talked less” item had lower endorsement among females compared to males matched on overall depressive symptoms.

\* Corresponding author. Tel.: (617) 278-0872; fax (617) 731-3843.

E-mail address: scole@rics.bwh.harvard.edu (Stephen R. Cole)

Finally, the effect of race on the measurement properties of the CES-D is unknown. Callahan and Wolinsky [15] reported difference in factor structure by racial group, but inference was limited due to missing data. To date, there has been no published report of item bias due to race in the CES-D.

We present results of item-level analyses of invariance to the exogenous variables age, gender, and race on all CES-D items among subjects from the New Haven component of the Established Populations for Epidemiologic Studies of the Elderly (EPESE).

## 2. Methods

### 2.1. Sample

The New Haven EPESE study ( $N = 2812$ ) was one of four NIA-funded studies that randomly sampled community-dwelling men and women 65 years of age or older in 1982 to identify predictors of morbidity, mortality, disability, and placement in long-term care facilities [17]. Subjects were included in the present analyses if they responded to all 20 CES-D items and information was available on all exogenous variables ( $N = 2340$ ). The remaining 2340 subjects were similar to the 472 removed from analysis with respect to gender, but were more likely to be black (20% vs. 13%,  $P < 0.01$ ) and were less likely to be older ( $P$  for trend  $< 0.01$ ).

### 2.2. Procedures

We examined the exogenous variables (1) age, which was collected in discrete categories and analyzed as a binary variable (less than 75 vs. 75 or older), (2) gender, and (3) race, which was collected and analyzed as black or white. We scored the CES-D as a summated rating scale, with each of the 20 items contributing between 0 and 3 points towards the summary scale, which has a range of 0–60 (items 4, 8, 12, & 16 are reverse scored). Because the CES-D summary score was skewed, we took the log transformation of the CES-D score. This log-transformed CES-D score was much less skewed, and is used for all analyses to provide better covariance adjustment.

Several methods are available for the analysis of item bias, including methods developed from item response theory [18], multisample confirmatory factor analysis, ordered correlation coefficients [19], and the Mantel-Hanszel (MH) adjustment method [18]. We believe the method outlined below, an extension of the MH method, is most appropriate for a medical and public health audience due to the use of proportional odds ratios.

The statistical formulation of the problem of differential item difficulty is:  $I_i \perp E_j \mid \theta$  [i.e., each studied item ( $I_i$ ) should be unrelated to each exogenous variable ( $E_j$ ), after conditioning on the latent variable ( $\theta$ )]. The MH odds ratio conditioned on the total scale score is a commonly used test of the statistical formulation above where both  $I_i$  and  $E_j$  are binary variables [18]. It is important to note the difference

between item impact and differential item difficulty (or item bias). Item impact is the crude (not conditioning on latent variable) difference in response to an item by level of the exogenous variable. Because individuals differ with respect to the latent variable, we expect item impact to vary. Differential item functioning, in contrast, is expected to be the same across the levels of the exogenous variable *once conditioned on the latent variable*.

We used a proportional odds regression model [20,21] to extend the commonly used MH odds ratio to the situation where the items,  $I_i$ , are ordered categorical items. An independent association between any studied item and exogenous variable (total scale-adjusted proportional odds ratio  $\neq 1.0$ ) provides evidence that response to the item is biased with respect to the exogenous variable. We explored the proportional odds assumption whenever the P-value from the proportional odds score test was  $< 0.10$ . To confirm our use of the proportional odds model, we calculated Spearman rank correlation coefficients between each item and each exogenous variable, partialing out the CES-D total score [19]. Because the findings were extremely similar by both methods, we do not present the partial rank correlation coefficients.

As the P-value confounds sample size with effect size, we retained all items that demonstrated a “relatively large” practically meaningful bias, which we considered to be a proportional odds ratio  $> 2.0$  or, conversely,  $< 0.5$ . A proportional odds ratio of 2.0 translates to those in the test group being at twice the odds of responding higher to the individual item than those in the control group, after being matched on overall depressive symptoms.

To test for item discrimination bias, for those items that demonstrated practically meaningful item difficulty bias, we included an interaction term between the exogenous variable and the total CES-D scale score. Additionally, we tested for differential *factor* functioning, or factor bias, by associating the factor score with each exogenous variable, while conditioning on the overall CES-D scale.

An important assumption of item bias analyses is that the scale measured by the various items represents a single underlying construct (i.e., unidimensionality). In the present case, the items on the CES-D are presumed to tap a single construct, namely depression. However, four subscales (depressed mood, low positive mood, somatic complaints, and interpersonal problems) have been reproduced on various samples to various degrees. Indeed, it is common practice to report the overall CES-D scale score, which is an implicit acceptance of a unidimensional scale. Moreover, Hertzog *et al.* [11] reported results from confirmatory factor analysis models that support the use of a total CES-D score.

Finally, we present a shortened revised version of the CES-D scale with items removed that function significantly differently by level of the exogenous variables age, gender, and race. While we do not propose that the full 20-item version is a gold standard measure of depression, comparison to the full CES-D scale is of interest, particularly to those

currently using the full scale. We present the Spearman rank correlation between the revised version and full scale. In addition, we present the sensitivity, specificity, and the percent area under the ROC curve using a score of 16 or greater on the full 20-item scale to define the threshold for diagnosis of clinical depression [1].

### 3. Results

Sample characteristics for the 2340 subjects were 58% female, 20% black, and 32% educated at or beyond 12th grade. The age distribution was 36% <70 years, 28% 70–74 years, 17% 75–79 years, 11% 80–84 years, and 8% >84 years. CES-D item responses are provided in Table 1. All item responses were skewed towards the Rarely or Never category. The average CES-D scale score was 8 (interquartile range: 2, 12), while the average log-transformed CES-D scale score was 1.7 (interquartile range: 1.1, 2.6). The internal consistency reliability of the CES-D, as measured by Cronbach's alpha, was 0.86.

#### 3.1. Item-level analysis

We observed 17 of the 20 CES-D items to be relatively free of item bias by age group, gender, and racial group. However, we did observe three occurrences of practically meaningful item-level bias (proportional odds ratio > 2.0 and Spearman rank correlation > 0.10; Table 2). The proportional odds of blacks responding higher on the item "people are unfriendly" were 2.29 times (95% confidence interval: 1.74, 3.02) that of whites matched on overall depressive symptoms. The mean score for the "people are un-

friendly" item, adjusted for overall depressive symptoms, was 0.37 (standard error 0.02) for blacks and 0.19 (standard error 0.01) for whites. The proportional odds of blacks responding higher to the item "people dislike me" were 2.96 times (95% confidence interval: 2.15, 4.07) that of whites matched on overall depressive symptoms. Neither of these differences in item difficulty varied by level of depressive symptoms (P for interaction = 0.53 for both tests). The mean score for the "people dislike me" item, adjusted for overall depressive symptoms, was 0.25 (standard error, 0.02) for blacks and 0.11 (standard error 0.01) for whites. As these two items combine to comprise the interpersonal problems factor of the CES-D, this item-level bias in favor of blacks matched on depressive symptoms reporting more interpersonal problems carries over as a positive factor-level bias, whereby the proportional odds of blacks responding higher on the interpersonal problems subscale were 2.72 times (95% confidence interval: 2.11, 3.51; Table 3) that of whites matched on overall depressive symptoms. Finally, the proportional odds of women responding higher on the item "crying spells" were 2.14 times (95% confidence interval: 1.60, 2.82) that of men matched on overall depressive symptoms. The mean score for the "crying spells" item, adjusted for overall depressive symptoms, was 0.23 (standard error 0.01) for women and 0.15 (standard error 0.02) for men. This difference in item difficulty by gender did not appear to vary by level of depressive symptoms (P for interaction = 0.86). There was no evidence of any item bias by age group in this sample of elders.

Table 1  
Item responses (N = 2340)

Item	Response (%)			
	Rarely/none of the time	Some of the time	Much of the time	Most/all of the time
1) Bothered by things	72	20	2	6
2) Poor appetite	78	13	2	6
3) Could not shake blues	79	14	3	4
4) As good as others	4	6	4	85
5) Trouble keeping mind on task	72	20	3	4
6) Felt depressed	65	26	4	5
7) Everything an effort	69	19	3	8
8) Felt hopeful	19	16	7	59
9) Life a failure	87	8	1	3
10) Felt fearful	80	15	1	3
11) Restless sleep	66	20	3	10
12) Happy	6	15	10	68
13) Talked less	78	13	3	5
14) Felt lonely	68	21	3	7
15) People unfriendly	85	10	1	3
16) Enjoyed life	5	10	7	77
17) Crying spells	86	10	2	2
18) Felt sad	67	26	3	3
19) People disliked me	90	7	1	2
20) Could not get going	76	17	3	4

Table 2  
Differential item functioning: proportional odds ratios between CES-D items and age, gender, and race, conditioned on total CES-D score<sup>a</sup> (N = 2340)

Item	75 or older	Female	Black race
1) Bothered by things	1.13	1.01	0.68
2) Poor appetite	0.94	0.93	1.40
3) Could not shake blues	1.09	1.25	0.79
4) As good as others	1.01	1.04	0.91
5) Trouble keeping mind on task	1.38	0.98	0.83
6) Felt depressed	0.87	1.08	0.77
7) Everything an effort	1.10	0.89	1.04
8) Felt hopeful	1.13	0.84	1.06
9) Life a failure	0.71	0.70	0.99
10) Felt fearful	0.97	1.35	1.12
11) Restless sleep	0.91	1.28	0.85
12) Happy	0.91	1.14	0.77
13) Talked less	0.99	0.70	0.99
14) Felt lonely	1.15	1.35	0.96
15) People unfriendly	0.74	0.74	<b>2.29<sup>b</sup></b>
16) Enjoyed life	1.12	1.07	0.68
17) Crying spells	1.26	<b>2.14</b>	0.62
18) Felt sad	0.89	1.30	0.93
19) People disliked me	0.73	0.85	<b>2.96</b>
20) Could not get going	0.87	1.24	1.04

<sup>a</sup> Log transformed CES-D score.

<sup>b</sup> Bold indicates Spearman Rank correlation exceeds 0.10.

Table 3

Differential factor functioning: proportional odds ratios between CES-D factors and by age, gender, and race, conditioned on total CES-D score<sup>a</sup> (N = 2340)

Factor	75 or older	Female	Black race
Somatic complaints <sup>b</sup>	1.05	0.96	0.86
Depressed mood <sup>c</sup>	0.96	1.37	0.84
Low positive mood <sup>d</sup>	1.02	0.98	0.84
Interpersonal problems <sup>e</sup>	0.74	0.79	<b>2.72<sup>f</sup></b>

<sup>a</sup>Log transformed CES-D score.

<sup>b</sup>Items # 1, 2, 5, 7, 11, 13, 20 (Table 1).

<sup>c</sup>Items # 3, 6, 9, 10, 14, 17, 18 (Table 1).

<sup>d</sup>Items # 4, 8, 12, 16 (Table 1).

<sup>e</sup>Items # 15 & 19 (Table 1).

<sup>f</sup>Bold indicates the Spearman Rank Correlation exceeded 0.10.

### 3.2. Revised 17-item CES-D scale

The reduced 17-item CES-D retained a high internal consistency reliability of 0.85. The 17-item version correlated 0.99 with the full 20-item version. Taking the standard cut-point of  $\geq 16$  points on the full 20-item scale as the threshold for diagnosis of clinical depression [1], the sensitivity and specificity of the reduced 17-item scale varied with the choice of cut-point, as shown in Table 4.

## 4. Discussion

We found three of 20 CES-D items to function differently among subgroups of gender and race. The two items that comprise the interpersonal problems factor of the CES-D were each biased towards higher endorsement among blacks, after matching on overall depressive symptom score. These item-level biases carried over as a factor-level bias, as these two items combine to comprise the interpersonal problems factor of the CES-D. We observed no evidence of item-level bias by age in this sample of elders.

The only prior work on item-level bias among adults using the CES-D is that of Strommel [16] among a sample of 1212, aged 18 to 88 years. In agreement with Strommel's findings, we found the crying item gender biased. However, we did not support Strommel's finding that the "talked less" item was gender biased.

We found no evidence of item bias by age, comparing those 75 and older to those 65 to 74 years of age. The absence of an age bias may be due to either no bias due to age or to the restricted age range in our sample of elders.

No previous work has addressed racial item bias in the CES-D. Our novel finding, that blacks were more likely to endorse higher levels of the two interpersonal problem items, may help to explain the low factor-level correlation between the interpersonal problems factor and the remaining CES-D factors (depressed mood, low positive affect, and somatic complaints) that has been repeatedly demonstrated [1,11]. We suspect that the interpersonal problem items are confounded with the perception of racial preju-

Table 4

Sensitivity and specificity for depression classification with reduced 17-item CES-D scale: gold standard of  $\geq 16$  on full 20-item CES-D scale (N = 2340)

Cut-point on 17-item CES-D scale	Sensitivity <sup>a</sup>	Specificity <sup>b</sup>	Area under ROC curve
16	88	100	93%
15	93	99	96%
14	97	97	97%
13	99	94	96%
12	100	91	96%

<sup>a</sup>Calculated as true positives/true positives + false negatives.

<sup>b</sup>Calculated as true negatives/true negatives + false positives.

dice, and therefore lack construct validity and undermine the applicability of the CES-D.

Unidimensionality is a crucial attribute of any multi-item measurement scale that presents a global score and is an essential assumption of item bias analyses. A simple hypothetical example will demonstrate the issue. Consider a 20-item measurement scale that purports to measure health-related quality of life by a single summary score. In truth, the 20 items elicit information on mental and physical components of health-related quality of life, which are related but distinct constructs. A global summary measure that reports a subject has a moderate score could be reflecting a moderate score on both constructs, a high mental and low physical component score, or a low mental and high physical component score. The lack of unidimensionality confuses the interpretation of the global summary score. Our data suggest that the two interpersonal problem items commingle the construct "perception of racial prejudice" with the construct that is intended to be measured, namely depression.

A second assumption of item bias analyses is that the variables upon which one compares the response to scaled items are exogenous. This means that these items *cannot be affected* by other variables under study. Age, gender, and race are exogenous variables. Endogenous are those that *can be affected* by variables under study. Education and annual income are endogenous variables. One can imagine a scenario in which a group defined by an exogenous variable (e.g., race) are ubiquitously chronically disadvantaged by lack of education or financial resources, such that this endogenous variable is nearly collinear with the exogenous variable defining the group. Under such a scenario, any association between a group defined by the exogenous variable and an item may be due, in part or fully, to the more proximate association between the endogenous variable, lack of education or resources, and the item under study. This did not appear to be the case in our data. Further adjustment for education and income level did not appreciably attenuate the associations between black race and "people are unfriendly" (adjusted proportional odds ratio = 2.01, 95% confidence interval: 1.49, 2.71), black race and "people dislike me" (adjusted proportional odds ratio = 2.73; 95% confidence interval: 1.92, 3.89) or female gender

and “crying spells” (Adjusted proportional odds ratio = 2.15; 95% confidence interval: 1.55, 2.99), after matching on depressive symptoms.

Item bias analyses often use a “purified subscale” method [18]. The purified subscale is defined as the total scale score minus biased items. The purified subscale is determined by successive passes through the data, whereby at each pass items found to be biased are removed from the purified subscale. Simulations can easily show that a purified subscale is necessary to estimate consistently the parameters for item difficulty. However, we question whether the choice of purified subscale is always clear in observed data, and whether this uncertainty in choosing the purified subscale might result in misspecification of the item bias parameters. Therefore, in our present study, we chose to condition on the full CES-D scale. However, we did explore the use of a purified subscale and found the item difficulty parameters were not appreciably altered by analysis using a purified subscale that consisted of the 17 unbiased CES-D scale items.

We purposefully selected an effect size cut-point (proportional odds ratio > 2.0) to define a meaningful level of item bias. We did so to take the emphasis off the P-value and focus on the measure of association. In total, we made 60 comparisons. If we had used the conservative Bonferroni correction method for estimating statistical significance, the P-value of a statistically significant association would have been <0.0008. All three of the biased items had accompanying P-values <0.0001.

In summary, our evidence indicates the CES-D is more valid in this diverse sample of elders after removal of the crying item and two interpersonal items. In our data, the reduced 17-item CES-D retains high internal consistency reliability, providing evidence that the removal of these items need not adversely affect the reliability of the CES-D. Establishing item invariance is of prime importance for drawing unbiased inferences in research using multi-item measurement scales [4,5]. Improvement in the measurement of latent constructs, such as negative emotions (depression, anger, and anxiety), may help to clarify ambiguous or inconsistent associations between these latent constructs and disease outcomes.

## Acknowledgment

Dr. Kawachi is supported by a Career Development Award from the National Heart, Lung, and Blood Institute.

## References

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

- [2] Thissen D, Steinberg L, Wainer H. Detection of differential item functioning using the parameters of item response theory. In: Holland PW, Wainer H, editors. *Differential Item Functioning*. Hillsdale, NJ: Erlbaum; 1993. pp. 67–114.
- [3] Nunnally JC, Bernstein IH. *Psychometric Theory*. Third edition. New York: McGraw-Hill; 1994.
- [4] Dean K, Holst E, Kreiner S, Schoenborn C, Wilson R. Measurement issues in research on social support and health. *J Epidemiol Community Health* 1994;48:201–6.
- [5] Dean K, Salem N. Detecting measurement confounding in epidemiological research: construct validity in scaling risk behaviours: based on a population sample in Minnesota, USA. *J Epidemiol Community Health* 1998;53(3):195–9.
- [6] Teresi J, Golden R, Cross P, Gurland B, Kleinman M, Wilder D. Item bias in cognitive screening measures: comparisons of elderly White, Afro-American, Hispanic and high and low education subgroups. *J Clin Epidemiol* 1995;48(4):473–83.
- [7] Groenvold M, Bjorner J, Klee M, Kreiner S. Test for item bias in a quality of life questionnaire. *J Clin Epidemiol* 1995;48(6):805–16.
- [8] Cole SR. Assessment of differential item functioning in the Perceived Stress Scale-10. *J Epidemiol Community Health* 1998;53:319–20.
- [9] Berkman LF, Berkman CS, Kasl S, Freeman DH, Leo L, Ostfeld AM, et al. Depressive symptoms in relation to physical health and functioning in the elderly. *Am J Epidemiol* 1986;124:372–88.
- [10] Liang J, Tran TV, Krause N, Markides KS. Generational differences in the structure of the CES-D scale in Mexican Americans. *J Gerontol* 1989;44(3):S110–20.
- [11] Hertzog C, Van Alstine J, Usala PD, Hultsch DF, et al. Measurement properties of the Center for Epidemiological Studies Depression Scale (CES-D) in older populations. *Psychol Assess* 1990;2(1):64–72.
- [12] Gatz M, Hurwicz M-L. Are old people more depressed? Cross-sectional data on Center for Epidemiological Studies Depression Scale factors. *Psychol Aging* 1990;5(2):284–90.
- [13] Kessler RC, Foster C, Webster PS, House JS. The relationship between age and depressive symptoms in two national surveys. *Psychol Aging* 1992;7(1):119–26.
- [14] Hays JC, Landerman LR, George LK, Flint EP, Koenig HG, Land KC, et al. Social correlates of the dimensions of depression in the elderly. *J Gerontol B Psychol Sci Soc Sci* 1998;53(1):31–9.
- [15] Callahan CM, Wolinsky FD. The effect of gender and race on the measurement properties of the CES-D in older adults. *Med Care* 1994;32:341–56.
- [16] Stommel M, Given BA, Given CW, Kalaian HA, Schulz R, McCorkle R. Gender bias in the measurement properties of the Center for Epidemiologic Studies Depression Scale (CES-D). *Psychiatry Res* 1993;49:239–50.
- [17] Cornoni-Huntley J, Ostfeld AM, Taylor JO, Wallace RB, Blazer D, Berkman LF, et al. Established populations for epidemiologic studies of the elderly: study design and methodology. *Aging* 1993;5(1):27–37.
- [18] Holland PW, Wainer H. *Differential Item Functioning*. New York: Erlbaum, 1993.
- [19] Stricker L. Identifying test items that perform differently in population subgroups: a partial correlation index. *Appl Psychol Meas* 1982;6:261–73.
- [20] Ananth C, Kleinbaum D. Regression models for ordinal responses: a review of methods and applications. *Int J Epidemiol* 1997;26(6):1323–33.
- [21] Scott SC, Goldberg MS, Mayo NE. Statistical assessment of ordinal outcomes in comparative studies. *J Clin Epidemiol* 1997;50(1):45–55.