Published in final edited form as:
Psychol Assess. 2011 June ; 23(2): 558–562. doi:10.1037/a0022484.

Psychometric Properties of Reverse-Scored Items on the CES-D in a Sample of Ethnically Diverse Older Adults

Mike Carlson[*], **Rand Wilcox**[†], **Chih-Ping Chou**[‡], **Megan Chang**[‡], **Frances Yang**[§], **Jeanine Blanchard**[*], **Abbey Marterella**[*], **Ann Kuo**[*], and **Florence Clark**[*]

[*] Division of Occupational Science and Occupational Therapy, School of Dentistry, University of Southern California, Los Angeles

[†] Department of Psychology, University of Southern California, Los Angeles

[‡] Institute for Health Promotion and Disease Prevention Research, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles

[§] Department of Medicine, Harvard Medical School; Institute for Aging Research, Hebrew SeniorLife, Boston

Abstract

Background—Reverse-scored items on assessment scales increase cognitive processing demands, and may therefore lead to measurement problems for older adult respondents.

Objective—To examine possible psychometric inadequacies of reverse-scored items on the Center for Epidemiologic Studies Depression Scale (CES-D) when used to assess ethnically diverse older adults.

Methods—Using baseline data from a gerontologic clinical trial (n=460), we tested the hypotheses that the reversed items on the CES-D: (a) are less reliable than non-reversed items, (b) disproportionately lead to intra-individually atypical responses that are psychometrically problematic, and (c) evidence improved measurement properties when an imputation procedure based on the scale mean is used to replace atypical responses.

Results—In general, the results supported the hypotheses. Relative to non-reversed CES-D items, the four reversed items were less internally consistent, were associated with lower item-scale correlations, and were more often answered atypically at an intra-individual level. Further,

Correspondence concerning this article should be addressed to Mike Carlson, Division of Occupational Science and Occupational Therapy, University of Southern California, 1540 Alcazar Street, CHP-133, Los Angeles, California 90089. Fax: (323) 442-1540; Phone: (909) 984-2856; mcarlson@usc.edu.
Mike Carlson, Ph.D., Division of Occupational Science and Occupational Therapy, School of Dentistry, University of Southern California, Los Angeles; Rand Wilcox, Ph.D., Department of Psychology, University of Southern California, Los Angeles; Chih-Ping Chou, Ph.D., Institute for Health Promotion and Disease Prevention Research, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles; Megan Chang, PhD, OTR/L, Institute for Health Promotion and Disease Prevention Research, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles; Frances Yang, Ph.D., Department of Medicine, Harvard Medical School; Institute for Aging Research, Hebrew SeniorLife, Boston; Jeanine Blanchard, M.A., OTR/L, Division of Occupational Science and Occupational Therapy, School of Dentistry, University of Southern California, Los Angeles; Abbey Marterella, M.S., OTR/L, Division of Occupational Science and Occupational Therapy, School of Dentistry, University of Southern California, Los Angeles; Ann Kuo, BSc, Division of Occupational Science and Occupational Therapy, School of Dentistry, University of Southern California, Los Angeles; and Florence Clark, Ph.D., OTR/L, Division of Occupational Science and Occupational Therapy, School of Dentistry, University of Southern California, Los Angeles.

the atypical responses were negatively correlated with responses to psychometrically sound non-reversed items that had similar content. The use of imputation to replace atypical responses enhanced the predictive validity of the set of reverse-scored items.

Conclusions—Among older adult respondents reverse-scored items are associated with measurement difficulties. It is recommended that appropriate correction procedures such as item re-administration or statistical imputation be applied to reduce the difficulties.

Keywords

CES-D; depression; reversed item format; older adults

Although the health of the older adult population can be promoted by high quality research on aging outcomes, such research is dependent on the availability of adequate assessments. Fortunately, standardized health-related outcome measures such as the Rand 36-Item Short Form Health Survey (SF-36) and the Center for Epidemiologic Studies Depression Scale (CES-D) perform reasonably well when used to assess a variety of older adult populations (Walters, Munro, & Brazier, 2001; Roberts, 1980; Radloff, 1977). However, despite their essential usefulness, some commonly used measures may contain particular types of items that create difficulties for elders, leading to measurement error.

Reverse-scored items (i.e., items for which a high score indicates the opposite of the construct being assessed, with a reverse-coding transformation applied) employ a structural format that may be especially problematic for older adult respondents. Such items force the respondent to (a) notice the altered direction of wording and (b) use the opposite end of the rating scale to produce a response that is consistent with the prior items. Although reverse-scored items serve a useful function by disrupting undesirable response sets such as acquiescence or disacquiescence, these benefits may be outweighed by the potential for methodologically induced bias (Pedhazur & Schmelkin, 1991; Nunnaly, 1967). Consistent with this possibility, across a variety of populations and assessments, it is common for reverse-scored items to cluster into a separate factor (Lyyra, Tormakangas, Read, Rantanen, & Berg, 2006; Long-Foley, Reed, Mutran, & DeVellis, 2002). Although these factors are often interpretable substantively, their content nonetheless co-varies with a reversed item format, raising the possibility that the loadings are at least partially methodologically based (Marsh, 1996; Marsh, 1986; Dunbar, Ford, Hunt, & Der, 2000).

Such a methodological effect could in part result from confusion that causes directionally opposite responding (e.g., among older adults who have reading difficulties). Due to the extreme deviation from the respondent's true standing on the item, such directionally based errors would be especially psychometrically disruptive.

In the present study, we examined the psychometric properties of reverse-scored items contained on the CES-D in a sample of ethnically diverse, primarily low income, community-dwelling older adults. We hypothesized that, due in part to directional response errors, reversed items are associated with: (a) a lower coefficient alpha and lower item-scale correlations; and (b) a greater percentage of individual item scores that are statistically deviant relative to a given elder's wider pattern of CES-D responses. Additionally, we hypothesized that (c) an individually applied correction procedure enhances psychometric outcomes for the set of such items.

METHODS

Participants

The participants were 460 men and women aged 60 years or more who enrolled in the University of Southern California Well Elderly 2 Study, an NIH-funded clinical trial of the effectiveness of a lifestyle intervention designed to reduce age-related declines among community-dwelling older adults (Jackson et al., 2009). Participants were recruited from 21 senior community centers, senior apartment complexes, or retirement communities in the Los Angeles area. At each recruitment site, prospective study participants met with a recruiter who explained the study requirements and obtained informed consent. Individuals unable to engage in the lifestyle intervention (e.g., due to severe cognitive deficits) or complete the study assessment battery with assistance were excluded.

Measures

The primary measure in the current study, the CES-D, is a widely used 20-item self-rating inventory that measures the frequency of depression symptoms experienced in the past week (Radloff, 1977). Items reflect depressed mood, feelings of guilt or worthlessness, perceptions of helplessness or hopelessness, and psychomotor/somatic symptoms. The frequency of each symptom during the past week is rated on a scale that ranges from 0 (rarely/none of the time, less than 1 day) to 3 (most or all of the time, 5–7 days). Scale items 4 (I felt that I was just as good as other people), 8 (I felt hopeful about the future), 12 (I was happy), and 16 (I enjoyed life) are reverse-scored and assess the absence of positive affect. CES-D test scores evidence adequate internal consistency and test-retest reliability, correlate with clinical judgments and self-report measures of depression, possess construct validity, and generate theoretically meaningful factors such as depressed affect, somatic symptoms, and diminished positive affect (Roberts, 1980; Radloff, 1977; Iwata & Buka, 2002).

The LSI-Z was used to measure life satisfaction, a construct which on theoretical grounds is expected to correlate negatively with depressive symptoms. This 13-item self-report measure, developed specifically for use with older adults, is internally consistent and possesses criterion-related validity (Wood, Wylie, & Sheafor, 1969).

Version 2 of the Short-Form 36-Item Health Survey (SF-36) was used to assess health-related quality of life (Ware, 2000). This widely accepted instrument includes eight domains that factor into physical and mental composites (Ware, 2000; Ware, Kosinski, & Keler, 1994). The two composite variables were included as validity criteria expected to correlate negatively with depressive symptoms.

Procedure

Data stemmed from self-administered questionnaires that were completed during the baseline phase of the clinical trial described above. In general, the assessments were administered to groups of 4 to 29 elders at the various recruitment sites. In addition to the CES-D, LSI-Z, and SF-36, participants responded to several additional scales and tests that were not analyzed in the present study (Jackson et al., 2009). Large print versions of the assessments were used to reduce eyestrain. A subset of the Hispanic participants (n=56) completed all study assessments in Spanish, using previously validated translated versions.

Data Analysis

The internal consistency of reversed versus non-reversed items was examined by comparing standardized coefficient alphas between the set of four reversed items and, to adjust for the influence of the number of items on coefficient alpha, the mean of the four sequential sets of four non-reversed items. Additionally, correlations between each CES-D item and the CES-

D total score (after removing the item) were calculated and the resulting correlation averages for reversed and non-reversed items were compared.

To assess the frequency of atypical responses to reversed vs. non-reversed items, we developed an algorithm for detecting *deviant* (i.e., outlying) individual item responses. In this procedure, the value (0, 1, 2, or 3) of each participant's item score most distant from the intra-individual mean was identified. If between 1 and 4 items shared this most distant value, the mean and standard deviation of the remaining set of 16 to 19 items were used to calculate a z-score for the item or items sharing the most distant value. (The value of the first quartile of the full sample's distribution of CES-D standard deviations was used as the denominator in calculating z-scores if a person's intra-individual standard deviation was zero for items other than those most distant from the mean.) All items sharing the most distant value were considered deviant if their (shared) z-score was 1.96 or greater in absolute value. A criterion value of 1.96 was used because it corresponds to a cut-off point (5%) that (a) is arguably sufficiently extreme to define deviant responses and (b) is the most commonly employed standard used to define a normal range (Colton, 1974). In this procedure, the reason for excluding the most distant items when calculating intra-individual means and standard deviations was to avoid masking effects based on outlier-induced distributional shifts. To allow for detection of multiple possible outliers on reverse-scored items, up to four deviant responses were permitted per participant. If five or more items shared the value most distant from the mean, the respondent was considered to have no deviant responses, based on (a) the degree of commonness (5/20 or more) of his or her most unusual response and (b) the joint contribution of reversed and non-reversed items to such responses, indicating at least a partial non-methodological underpinning.1

To describe overall differences in deviant responses between reversed and non-reversed items, for each item type we divided the number of deviant responses by the total number of responses across all research participants. To statistically test for item-type differences, for each participant we determined whether the proportion of deviant responses on reversed items was less than, was equal to, or exceeded the proportion of deviant responses on non-reversed items. Excluding instances that were equal, we then conducted a sign test, based on the results of Pratt (1968), of the null hypothesis of equality in the proportions of deviant responses for the two item types. This test was two-tailed and conducted at the .05 level of significance.

To control for item content in assessing the psychometric properties of deviant responses to reversed items, we conducted analyses involving reversed and non-reversed *pair-mates* (i.e., pairs of items with highly similar content, despite having a varying directional format). These analyses were conducted separately for the following two pair-mates which could be identified on the CES-D: (a) items 12 (I was happy) and 18 (I felt sad) and (b) items 16 (I enjoyed life) and 6 (I was depressed). For each pair, considering only the subset of older adults who produced a deviant response on the reversed item, we calculated a Pearson correlation coefficient between the two pair-mates, and also correlated each pair-mate with the corrected total CES-D score. These correlations were tested at the .05 significance level using a two-sided alternative.

To examine the effect of imputation for deviant responses to reverse-scored items, we replaced each such deviant response with the respondent's mean for all other CES-D items

[1]The sensitivity of the results to alternate outlier detection methods was examined by using two further algorithms: [a] use of the above strategy in connection with a more stringent z-score cut-point of 2.33, and [b] comparison of item scores against individualized medians through a univariate version of the minimum volume ellipsoid (MVE) estimator (Rousseeuw & Leroy, 1987). Because both of these procedures produced results similar to the main outlier detection method, they are not further discussed in this article.

(i.e., all non-reversed items and non-deviant reversed items). For comparison purposes, we then calculated basic statistics and validity correlations, across the entire sample, for the following five CES-D variants: (a) the intra-individual mean of all deviant responses to reverse-scored items (excluding participants with no such responses); (b) the sum of all reverse-scored items, without imputations; (c) the sum of all reverse-scored items, using imputations; (d) the sum of all CES-D items, without imputations; and (e) the sum of all CES-D items, using imputations. To assess predictive validity, we calculated Pearson correlation coefficients between each of the above variants and the SF-36 mental composite, SF-36 physical composite, LSI-Z, and uncorrected CES-D total. We tested each correlation at the 0.05 significance level using a one-tailed test in expectation of a theoretically congruent result.

RESULTS

The mean age of the 460 participants was 74.9 +/− 7.7 years; 65.9% were female; 62.6% were non-White (32.4% African American, 20.0% Hispanic/Latino, 3.9% Asian American, 6.3% Other/Refused); 49% had an education level of high school or less; and 53.5% had an annual income under $12,000.

The standardized coefficient alpha for the set of reversed items was lower than the average coefficient alpha for same-sized sets of non-reversed items (.67 versus .73). Consistent with this pattern, the mean item-scale correlation was lower for reversed than for non-reversed items (.45 versus .59).

The proportions of deviant responses for non-reversed and reversed items, respectively, were .054 (396/ [16*460]) and .143 (264/ [4*460]). Based on the sign test, the rate of deviant responding was significantly higher for reversed than for non-reversed items ($p < .001$). Of the 264 deviant responses to reversed items, 227 (86.0%) reflected the high end of the scale (a response of 2 or 3), indicating that the bulk of such responses were made by relatively non-depressed individuals, and tended to inflate depression scores. The percentages of deviant responses for the four individual reversed items were .15 (CES-D item 4), .24 (CES-D item 8), .11 (CES-D item 12), and .08 (CES-D item 16). Apart from the increase in deviant responses to the second reverse-scored item (CES-D item 8), the overall pattern reflects a decline in deviant responding to successively encountered items (5 of the 6 possible comparisons are consistent with this trend). Among the 460 participants, 62.4 %, 24.3%, 8.3%, 3.3%, and 1.7% had 0, 1, 2, 3, or 4 deviant responses, respectively, to reversed items.

In the analyses of pair-mates, negative correlations were obtained between the reversed items that were answered atypically and their non-reversed counterparts with similar content ($r = -.33$, $p = .021$ for items 12 and 18; $r = -.58$, $p < .001$ for item 16 and 6). Also, in both cases the non-reversed pair-mate was positively correlated with the corrected CES-D sum ($r = .65$, $p < .0001$ for item 18 and $r = .92$, $p < .0001$ for item 6), whereas the reversed pair-mate was negatively correlated with the corrected CES-D sum ($r = -.43$, $p < .002$ for item 12 and $r = -.59$, $p < .0001$ for item 16).

Psychometric properties of non-imputed and imputed CES-D variants are presented in Table 1. The per participant mean of deviant responses to reversed items correlated in a theoretically opposite manner with the SF-36 mental composite (r = .32), LSI-Z (r = .24), and CES-D Total (r = −.53). Considering the sum of all reversed items, the use of imputation reduced the mean per participant sum by one unit (from 3.41 to 2.41 points) and increased meaningfully the degree of theoretically predicated association with the three external validity criteria (mean r-square = .27 with vs. .16 without imputation) and the CES-

D total score (r-square = .72 with vs. .37 without imputation). With respect to CES-D total scores, the imputation procedure increased coefficient alpha from .90 to .93 and decreased the overall mean by one unit. However, the validity correlations for the CES-D total scores were not significantly affected by the use of imputation.

DISCUSSION

Overall, the results confirm that CES-D reverse-scored items are associated with measurement problems among older adults. Consistent with earlier research on a variety of populations, such items were as a set less internally consistent than non-reversed items and were more weakly associated with total scale scores (Long-Foley et al., 2002; Lee et al., 2008). Further, at the intra-individual level, reversed items were answered atypically 2.65 times more often than non-reversed items. The extremity of these deviant responses, in conjunction with their inverse association with psychometrically sound non-reversed items of like content, their negative correlation with the remainder of the CES-D, their theoretically “opposite” validity correlations with the LSI-Z and SF-36, and their tendency to occur more frequently for items encountered earlier on the assessment, supports the notion that the poorer psychometric performance of reversed items is at least in part due to directionally based information processing errors.

Depending on the context of usage, the effect of deviant responses to reversed CES-D items could be quite pronounced. For example, over 13% of the participants may have scored approximately 5 to 10 units higher or lower on the CES-D due to such atypical responses. Because such responses tended to gravitate to the high end of the scale, in comparison with mean-imputed scores they potentially inflated the sample-wide CES-D mean by nearly 8% (from 12.73 to 13.73) and positive affect mean by 41% (from 2.41 to 3.41, based on the sum of the four reversed items). In this regard, it is noteworthy that, using a cutoff score of 16 or greater, the number of elders who met the criterion for depression was reduced by approximately 10% (147 versus 162) when imputations were made for deviant responses to reversals.

Prior to drawing practical recommendations for altering the CES-D scoring procedure, it is important to consider a key limitation of the study, namely, that the results were to some degree affected by the specific content of particular items in addition to their reversed versus non-reversed status. For example, the second reversed item, which pertains to hope, may have often been answered atypically, though accurately, due to its focus on perceptions regarding the future that are somewhat independent of one’s immediate degree of depression. Thus, as a set the deviant responses are likely to have reflected a composite involving at least three ingredients: (a) errors based on directionally based mistakes (as well as any other types of errors, such as those stemming from culturally linked misunderstandings of item content), (b) veridical responses that inaccurately reflect depressive symptomatology, and (c) veridical responses that do in fact capture depressive symptomatology (as when a lack of hope is linked to a depressed state, despite low depression scores on the majority of CES-D items).

However, the above inability to disentangle the reasons underlying atypical responses does not imply that a correction procedure is unwarranted. In particular, when the combined influence of information processing errors and depression-independent veridical responses outweighs that of outlying depression-relevant responses, then correction is likely to produce a desirable tradeoff which fosters improved assessment. For example, a crude approximation might suggest that on theoretical grounds a psychometrically sound item (i.e., an item that evidences depression when answered properly) should produce a deviant response, as defined herein, approximately 5% of the time. If the percentage of deviant

responses exceeds 10%, then it is arguable that factors other than depressive symptomatology are equally or more likely to have produced a given scale-discrepant score, and imputation for intra-individually deviant responses should be considered as a means of fostering increased accuracy. For the reversed items in the current study (which were deviant 14.3% of the time), the improved associations with validity criteria following imputation with the scale mean support this contention. The results also support the use of non-reversed pair-mate scores to replace deviant responses to items 12 and 16. In analyzing research-based sample data, such statistical imputation represents a feasible correction procedure that can be performed any time following testing.

In clinical diagnostic contexts, or in research contexts when feasible, consideration should be given to solutions that rectify intra-individually deviant responses at (or closely following) the time of questionnaire administration. In this vein, respondents could be subjected to further questioning or be instructed to reconsider items that produce deviant answers (based on the application of an outlier detection program).

The phenomenon of spuriously induced intra-individual deviant responses is likely to characterize the measurement process more generally across different respondent populations, assessments, and item types. For example, 4 of the 16 non-reversed CES-D items were answered atypically more than 10% of the time in the current study. Future research should be conducted to document the degree of generality of this problem across varying assessment contexts and examine the relative utility of solutions such as re-questioning or imputation.

Acknowledgments

Supported by: National Institute on Aging, R01 AG021108

References

Colton, T. Statistics in medicine. Boston: Little, Brown, & Co; 1974.

Dunbar M, Ford G, Hunt K, Der G. Question wording effects in the assessment of global self-esteem. European Journal of Psychological Assessment. 2000; 16:13–19.

Iwata N, Buka S. Race/ethnicity and depressive symptoms: A cross-cultural/ethnic comparison among university students in East Asia, North and South America. Social Science and Medicine. 2002; 55:2243–2252.10.1016/S0277-9536(02)00003-5 [PubMed: 12409137]

Jackson J, Mandel D, Blanchard J, Carlson M, Cherry B, Azen S, Clark F. Confronting challenges in intervention research with ethnically diverse older adults: The USC Well Elderly II trial. Clinical Trials. 2009; 6:90–101.10.1177/1740774508101191 [PubMed: 19254939]

Lee SW, Stewart SM, Byrne BM, Wong JP, Ho SY, Lee PW, Lam TH. Factor structure of the Center for Epidemiological Studies Depression Scale in Hong Kong adolescents. Journal of Personality Assessment. 2008; 90:175–184. [PubMed: 18444112]

Long-Foley K, Reed PS, Mutran EJ, DeVellis RF. Measurement adequacy of the CES-D among a sample of older African–Americans. Psychiatry Research. 2002; 109:61–69. [PubMed: 11850052]

Lyyra TM, Tormakangas TM, Read S, Rantanen T, Berg S. Satisfaction with present life predicts survival in octogenarians. Journal of Gerontology Series B: Psychological Sciences and Social Sciences. 2006; 61:P319–26.

Marsh HW. The bias of negatively worded items in rating scales for young children: A cognitive-developmental phenomenon. Developmental Psychology. 1986; 22:37–49.

Marsh HW. Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? Journal of Personality and Social Psychology. 1996; 70:810–819. [PubMed: 8636900]

Nunnaly, J. Psychometric Theory. New York: McGraw-Hill; 1967.

Pedhazur, EJ.; Schmelkin, LP. Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Erlbaum; 1991.

Pratt JW. A normal approximation for binomial, f, beta, and other common, related tail probabilities, II. Journal of the American Statistical Association. 1968; 63:1457–1483.

Radloff LS. The CES-D Scale: A self-report depression scale for research in the general population. Applied Psychological Measurement. 1977; 1:385–401.

Roberts RE. Reliability of the CES-D Scale in different ethnic contexts. Psychiatry Research. 1980; 2:125–134. [PubMed: 6932058]

Rousseeuw, PJ.; Leroy, AM. Robust regression and outlier detection. New York: Wiley; 1987.

Walters SJ, Munro JF, Brazier JE. Using the SF-36 with older adults: a cross-sectional community-based survey. Age and Ageing. 2001; 30:337–343. [PubMed: 11509313]

Ware JE. SF-36 health survey update. Spine. 2000; 25:3130–3139. [PubMed: 11124729]

Ware, J.; Kosinski, M.; Keler, SD. SF-36 Physical and mental health summary scales: A user's manual. Boston: The Health Institute; 1994.

Wood V, Wylie ML, Sheafor B. An analysis of a short self-report measure of life satisfaction: Correlation with rater judgments. Journal of Gerontology. 1969; 24:465–469. [PubMed: 5362358]

Table 1

Descriptive Statistics and Validity Correlations for Selected CES-D Variants.

	Descriptive Statistics				Validity Criteria			
CES-D Variant	**n**	$\bar{x}$	**SD**	**Coefficient alpha**	**SF-36 Mental Composite**	**SF-36 Physical Composite**	**LSI-Z Total Score**	**CES-D Total Score**
Reversed Items Answered Deviantly: No Imputations (1–4 items)	173	2.33	.89	-	.32*	−.09	.24*	−.53*
Sum of Reversed Items: No Imputations (4 items)	460	3.41	2.95	.66	−.51*	−.19*	−.44*	.61*
Sum of Reversed Items: Imputations (4 items)	460	2.41	2.91	.87	−.66*	−.22*	−.58*	.85*
Sum of All CES-D Items: No Imputations (20 items)	460	13.73	10.91	.90	−.75*	−.23*	−.59*	(1.00)
Sum of All CES-D Items: Imputations (20 items)	460	12.73	12.75	.93	−.73*	−.22*	−.58*	.98*

Note.

*p < .05. CES-D indicates Center for Epidemiology Studies Depression Scale; SF-36, Short-Form 36-Item Health Survey; LSI-Z, Life Satisfaction Index-Z.