

Using a depression-scale item to measure loneliness: low validity and tautological inflation of the loneliness–depression association

Ziyan Wang *

School of Public Health, Hebei Medical University, Shijiazhuang 050017, China

* Corresponding Author Email: 13427533552@qq.com

Abstract. In several large ageing cohorts, loneliness is measured with the single "felt lonely" item of the Center for Epidemiologic Studies Depression (CES-D) scale, while the CES-D total score serves as the depression outcome, so that one item stands on both sides of the association. Using the 2018 wave of the Health and Retirement Study (HRS) leave-behind subsample (N = 5,083, survey-weighted, Rao–Wu rescaled cluster bootstrap with 1,000 replicates), two properties of this practice were quantified in the same respondents. The adjusted coefficient of the item was 3.2408 against the eight-item CES-D and 2.2408 against the same scale with the item removed; the difference was 1.0000 in every bootstrap replicate, an algebraic consequence of the item being one of the components of the total. In the main specification this definitional component accounts for 30.9% (95% confidence interval 29.3–33.0) of the contaminated coefficient, and for between 27.6% and 30.9% across twelve alternative specifications. Against the three-item University of California, Los Angeles (UCLA) loneliness scale the item had a sensitivity of 37.7% and a kappa of 0.363, missing 62.3% of respondents classified as lonely; this criterion-validity comparison independently replicates and extends the validation reported by Yu et al. (2025) in the same cohort. In a mediation specification the proportion mediated was 1.33 times its value under the clean specification, with an interval of 0.63–2.29 that includes unity. The item is neither an adequate measure of loneliness nor independent of the outcome it is used to predict.

Keywords: loneliness; depressive symptoms; CES-D; measurement validity; criterion contamination.

1. Introduction

Loneliness is among the most consistently reported psychosocial correlates of depressive symptoms in later life. A recent meta-analytic structural equation model synthesising 53 studies and 40,929 participants placed the standardised loneliness–depression association at 0.28 [1]. Estimates of this magnitude, replicated across countries and designs, have made loneliness a standard construct in ageing cohort studies, and most major cohorts now field some measure of it.

The instrument used, however, is frequently not a loneliness scale. In several large ageing cohorts loneliness is operationalised with a single question — "I felt lonely" — drawn from the Center for Epidemiologic Studies Depression (CES-D) scale [2]. This is the practice in the Health and Retirement Study (HRS) [4], the Survey of Health, Ageing and Retirement in Europe [5], the Canadian Longitudinal Study on Aging [6] and the China Health and Retirement Longitudinal Study [7]. These four studies address widely different outcomes, from memory decline to frailty; what they share is the choice of exposure measure. The total score of the same instrument is also, however, the standard measure of depressive symptoms in these cohorts, and the eight-item version administered in the HRS retains the loneliness item among its eight components [3]. Whenever that exposure is combined with that outcome, one item stands simultaneously on the exposure side and inside the outcome. In psychometrics this configuration is criterion contamination, a problem examined for the somatic items of depression scales in chronic-disease populations [8]; no application of it to the loneliness item was found in the present search.

That the overlap is recognised is evident from studies that removed the item before analysis, though the removals were made in different designs and for different purposes. Lee et al., whose exposure was a separate loneliness scale, removed the item from the CES-D outcome because the overlap

"might inflate associations", and reported the resulting change in internal consistency (Cronbach's alpha 0.78 to 0.76) but not the change in the association itself [9]. Salinas et al. removed it from a CES-D score entered as a covariate, in an analysis whose exposure was the item and whose outcome was incident dementia [10]. In neither study was the single item regressed on a depression total containing it, so neither could report — and neither reports — how large the resulting inflation would be.

The 2018 HRS wave administered both instruments to the same respondents: the contaminated single item, and the three-item UCLA loneliness scale developed for large surveys and validated against the revised UCLA scale [11]. This design permits both questions to be answered in one sample [12]. Three contributions follow. First, the contaminated coefficient is decomposed. It contains a component of exactly one that is definitional rather than empirical: this follows from least-squares algebra and holds for any covariate set, weighting or sample, while the share of the coefficient that the component represents depends on the specification and is 30.9% in the main analysis. Second, the finding of Yu et al. [4] that the single item has low sensitivity against the three-item UCLA scale is independently replicated in a later wave with an expanded set of agreement statistics, and is shown to hold across dichotomisation thresholds from ≥ 4 to ≥ 8 . Third, the two problems are shown to be properties of the same item, and their propagation into a commonly estimated mediation specification is traced, with the resulting inflation reported together with its confidence interval.

2. Methods

2.1. Data and study sample

The HRS is a nationally representative longitudinal study of United States residents aged 50 years and older [12]. Three public-release files were merged on the household and person identifiers: the RAND HRS Longitudinal File 2022 [15], [16], the 2018 HRS Core leave-behind psychosocial questionnaire [17], and the Cross-Wave Tracker file [18]. The 2018 wave corresponds to RAND wave 14 and to the HRS Core prefix Q.

The leave-behind psychosocial questionnaire is administered to a rotating half of the sample, and the 2018 wave was fielded to subsample A; the file therefore contains 17,146 rows of which the majority are missing on all leave-behind items. Eligibility was defined by a positive leave-behind respondent weight, which the HRS generates only for eligible respondents who returned the questionnaire. Of 46,766 respondents in the tracker file, 5,491 had a positive weight; requiring complete data on the three UCLA items reduced this to 5,331, on the CES-D total and the loneliness item to 5,330, on the five participation items to 5,121, and on all covariates to the analytic sample of $N = 5,083$.

The 2018 wave was chosen because it is the most recent pre-pandemic wave with a complete field period. The 2020 wave replaced in-person interviewing with a mail and web protocol and added a coronavirus module, which makes it a poor basis for a measurement comparison.

2.2. Measures

The contaminated loneliness measure is the CES-D loneliness item as recorded in the RAND file. In the HRS the CES-D is an eight-item version with a yes/no response format [3], so this measure is binary. The contaminated depression outcome is the eight-item CES-D total score (range 0–8) and the de-contaminated outcome is the same score with the loneliness item removed (range 0–7).

The reconstruction of the total score is not a matter of addition. Two of the eight items, "was happy" and "enjoyed life", are stored in the RAND file in their original positive direction and must be reversed before summation. Summing the eight stored values reproduces the released total for only 8.01% of respondents, whereas reversing these two items first reproduces it for 99.84%. This identity, $\text{CES-D-8} \equiv \text{CES-D-7} + \text{loneliness item}$, is the premise of the decomposition and was verified before any analysis.

The clean loneliness measure is the three-item UCLA loneliness scale [11], administered in the leave-behind questionnaire as how often the respondent feels a lack of companionship, feels left out, and feels isolated from others. In the HRS data file these items are coded in the direction opposite to the published scale, with higher values indicating lower loneliness; each was therefore reverse-coded, giving a summed score from 3 to 9 in which higher values indicate greater loneliness. The correctness of this direction is established by the data rather than assumed: the correlation between the reverse-coded scale and the CES-D loneliness item is +0.436, and a reversal error would have produced a negative sign. For the criterion-validity analysis the scale was dichotomised at ≥ 6 , the threshold used by Yu et al. in the same cohort [4]; thresholds from ≥ 4 to ≥ 8 are reported alongside it.

Social participation is used here as an illustrative exposure because the participation–loneliness–depression pathway is among the most frequently estimated in ageing cohorts. The substantive association is not the target of this study; the estimand concerns how the measurement of loneliness changes the estimated association. Participation was measured by the frequency of five social activities in the leave-behind questionnaire — voluntary work with young people, charity work, attending a sports, social or other club, attending a non-religious organisation, and taking part in a community arts group — each recorded on a seven-point scale on which higher values denote lower frequency. Items were reverse-coded and summed, giving a score from 5 to 35 in which higher values denote greater participation.

Covariates were age in years, years of education, self-rated health (1 = excellent to 5 = poor), sex, race (White versus other) and marital status (married or partnered versus other). All estimates were weighted by the leave-behind respondent weight, with the HRS stratum and primary sampling unit identifiers used for variance estimation (80 strata, 159 primary sampling units).

The participation measure, the covariate set and the complete-case rule were fixed before the results were inspected. Alternative operationalisations are not reported selectively; they are presented together in the specification curve.

2.3. Statistical analysis

Associations were estimated by weighted least squares. The primary estimand is at the level of the regression coefficient: b_1 , the adjusted coefficient of the loneliness item on the eight-item CES-D; b_2 , its coefficient on the seven-item score; their ratio; and the definitional share $1/b_1$. The proportion mediated was treated as a secondary quantity, because it is a ratio of ratios whose instability under weak denominators is a known limitation and is unrelated to the argument made here.

Criterion validity was assessed against the dichotomised UCLA scale using sensitivity, specificity, predictive values, Cohen's kappa, the area under the receiver operating characteristic curve, the Pearson correlation and the raw agreement rate. The raw agreement rate is reported because it is the statistic on which a recent comparison in another cohort based its conclusion [6], not because it is an adequate criterion.

Confidence intervals come from a Rao–Wu rescaled cluster bootstrap with 1,000 replicates and percentile intervals. The HRS is a paired-primary-sampling-unit design: 79 of its 80 strata contain two primary sampling units and one contains a single unit. Resampling all n_h units with replacement within each stratum, which is the most common form of cluster bootstrap, has an expected variance of $(n_h - 1)/n_h$ times the design variance; under a paired design this halves it, and every standard error reported here would have been understated by a factor of approximately $\sqrt{0.5}$. In each replicate, therefore, $n_h - 1$ units were drawn with replacement within each stratum and the weights of the drawn units were multiplied by $n_h/(n_h - 1)$, with unselected units receiving zero weight; the single-unit stratum was carried through unchanged. The standard error of b_1 obtained in this way agrees to within 1.2% with a Taylor-linearisation cluster-robust standard error for the same model.

A specification curve was estimated across 336 combinations of loneliness measure, depression outcome, covariate set, missing-data handling, participation measure and weighting; the estimand

there is the standardised coefficient, because the binary item and the continuous scale are not comparable on their raw scales. Standardised coefficients for a binary measure and for a continuous scale are not, however, equivalent quantities: the standard deviation of a binary variable is a function of its prevalence, and a single item and a three-item scale differ in reliability. The specification curve is therefore read as a comparison of what each analytic choice produces, and not as an estimate of one underlying quantity on a common scale.

The data are cross-sectional, and every quantity reported in this paper is an association rather than an effect.

3. Results

3.1. Sample characteristics

Table 1 reports the weighted characteristics of the analytic sample; the weights sum to approximately 97.3 million United States adults aged 50 and over.

Table 1. Weighted characteristics of the analytic sample (HRS 2018 leave-behind subsample, N = 5,083).

Characteristic	Value
Age, years, mean (SD)	66.1 (9.7)
Female, %	53.7
White, %	81.0
Married or partnered, %	66.6
Education, years, mean (SD)	13.5 (2.8)
Self-rated health (1 = excellent to 5 = poor), mean (SD)	2.8 (1.0)
CES-D-8 score, mean (SD)	1.3 (1.9)
UCLA 3-item loneliness score, mean (SD)	4.5 (1.7)

3.2. The single item is tautologically bound to the outcome

Adjusted for social participation and all covariates, the coefficient of the loneliness item was 3.2408 (95% CI 3.0272–3.4097) when the outcome was the eight-item CES-D and 2.2408 (95% CI 2.0272–2.4097) when the outcome was the seven-item score. The difference was 1.000000000000 in every one of the 1,000 bootstrap replicates: the minimum and the maximum of its bootstrap distribution agree to twelve decimal places and its standard deviation is zero. The two distributions are shown in Fig. 1.

This is not an empirical regularity but an algebraic identity, and it can be stated without reference to any data. Write the eight-item score as $Y_8 = Y_7 + M$, where M is the loneliness item, and let M be one of the columns of the design matrix. Least squares is linear in the outcome, so the coefficient vector for Y_8 equals the coefficient vector for Y_7 plus the coefficient vector obtained by regressing M on the same design matrix. Because M is itself a column of that matrix, the second of these is the unit vector selecting M . The coefficient of M therefore rises by exactly one and every other coefficient is unchanged — irrespective of covariates, survey weighting, sample size, and indeed of the dataset. The argument uses only the linearity of least squares in the outcome, so it holds for any linear additive specification; it does not extend to models in which the item enters non-linearly or interactively, or to non-linear link functions. The bootstrap in Fig. 1 estimates the sampling variability of the two coefficients; it does not estimate the separation between them, which the algebra fixes.

In this sample the consequence is that 30.9% (95% CI 29.3–33.0) of the contaminated coefficient is definitional, and that the coefficient is 1.4463 times (95% CI 1.4150–1.4933) the same coefficient estimated against the de-contaminated outcome.

The component and the share it represents are different quantities and behave differently. The definitional component is exactly 1.000000000000 under all twelve combinations of covariate set, participation measure and weighting that were examined, as the algebra requires. The share is $1/b_1$, whose denominator is the contaminated coefficient itself; it therefore does depend on the specification, and it ranged from 27.6% to 30.9% across those same twelve combinations. Invariance is a property of the component, not of the share.

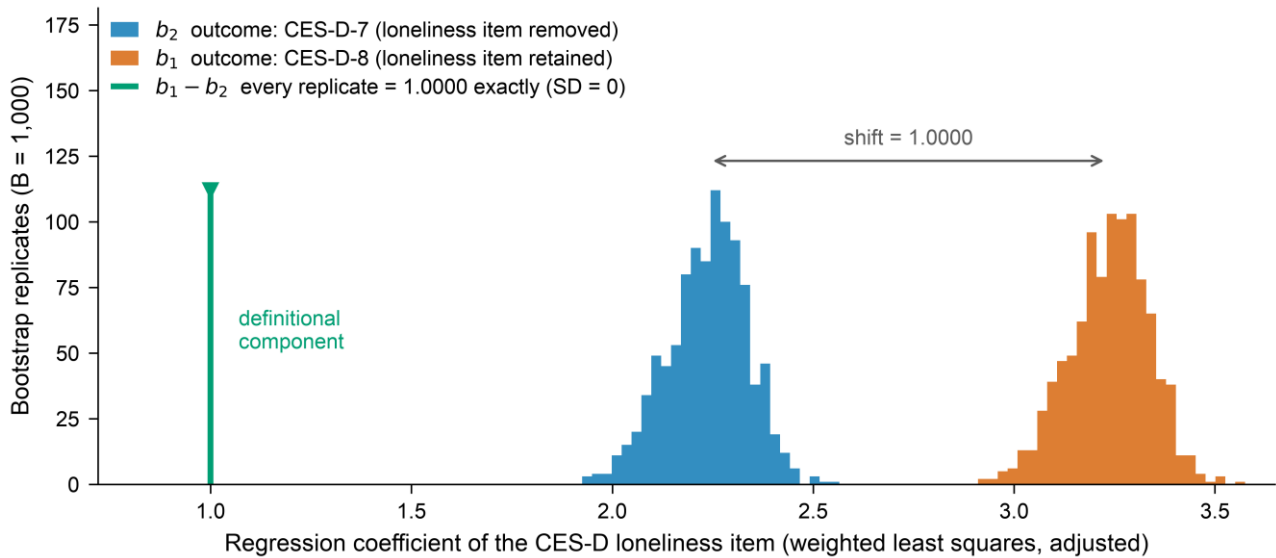


Fig. 1 Bootstrap distributions of the loneliness coefficient under the contaminated and de-contaminated outcomes (1,000 replicates). What is estimated is the sampling variability of each coefficient; the separation between them is not estimated but fixed by the algebra at exactly 1.0000, so their difference is degenerate and has zero sampling variance

3.3. The single item is a poor measure of loneliness

Consistent with Yu et al. [4], who reported sensitivities of 0.30 in men and 0.39 in women against the same reference standard and threshold in the 2006 HRS wave ($n = 7,445$), the single item performed poorly as a measure of loneliness in 2018 (Table 2). The three-item UCLA scale classified 25.4% of respondents as lonely ($n = 1,291$), the single item 14.5% ($n = 737$), a detection rate 0.57 times that of the scale. Sensitivity was 37.7% and specificity 93.4%; the item missed 804 respondents classified as lonely by the scale, or 62.3% of that group. Agreement beyond chance was poor (kappa 0.363); the correlation with the continuous scale was 0.436 and the area under the curve for the item against the dichotomised scale was 0.656. Survey weighting did not change this picture (weighted sensitivity 36.7%, specificity 93.6%).

The result does not depend on where the reference scale is dichotomised. Across thresholds from ≥ 4 to ≥ 8 sensitivity never exceeded 53.5% and kappa never exceeded 0.373. The threshold used in the main analysis is not the one least favourable to the single item: kappa at ≥ 6 (0.363) is close to its maximum across the range (0.373 at ≥ 7), which rules out a threshold chosen to disadvantage the item. That threshold was in any case taken from the earlier validation in this cohort rather than selected here [4].

Table 2. Agreement between the CES-D single item and the UCLA three-item scale as measures of loneliness, and robustness to the dichotomisation threshold (N = 5,083).

Panel A: classification at the ≥ 6 threshold			UCLA: not lonely		UCLA: lonely
Single item = 0			3,542		804
Single item = 1			250		487
Agreement statistic					Value
Sensitivity, %					37.7
Specificity, %					93.4
Positive predictive value, %					66.1
Negative predictive value, %					81.5
Cohen's kappa					0.363
Raw agreement, %					79.3
Pearson correlation with the continuous scale					0.436
Area under the curve, item against the dichotomised scale					0.656
Area under the curve, continuous scale against the item					0.813
Panel B: threshold	Classified lonely, %	Sensitivity, %	Specificity, %	Kappa	Missed, n (%)
≥ 4	56.9	23.5	97.4	0.187	2,211 (76.5)
≥ 5	39.6	29.8	95.5	0.284	1,412 (70.2)
≥ 6 (main analysis)	25.4	37.7	93.4	0.363	804 (62.3)
≥ 7	11.3	51.6	90.2	0.373	279 (48.4)
≥ 8	6.0	53.5	88.0	0.248	141 (46.5)

The two areas under the curve answer different questions and are not interchangeable. The discriminative performance of the single item, which is the quantity of interest, is 0.656; the value of 0.813 is obtained in the reverse direction, using the continuous scale to discriminate who endorsed the item, and is reported only to show that the scale carries information about the item that the item does not carry about the scale.

3.4. Consequences for a commonly estimated mediation specification

Table 3 reports the participation–loneliness–depression pathway under the four combinations of loneliness measure and depression outcome. The proportion mediated was 39.7% when both the contaminated measure and the contaminated outcome were used and 30.0% under the clean specification, a ratio of 1.33. The bootstrap interval for this ratio was 0.63–2.29 and includes unity, so the inflation of the proportion mediated is not statistically significant. The width of that interval is explained by the table: both the total association c and the a path of the binary item are weak, and the proportion mediated is a ratio of ratios whose instability under a small denominator is a known property of the estimator rather than a feature of these data. The primary estimand of this study is for that reason the regression coefficient decomposition and not the proportion mediated.

The a paths of the two measures are not directly comparable in magnitude (-0.0050 for the binary item versus -0.0347 for the 3–9 scale) because the measures differ in scale; the proportion mediated is invariant to linear rescaling of the mediator, so the four cells remain comparable.

Table 3. Loneliness–depression association under four measurement specifications, with Rao–Wu rescaled bootstrap 95% confidence intervals (1,000 replicates).

Specification	a	b	c	a·b	c'	Proportion mediated, % (95% CI)
Contaminated item $\times$ CES-D-8	−0.0050	3.2408	−0.0406	−0.0161	−0.0245	39.7 (20.7–56.8)
Contaminated item $\times$ CES-D-7	−0.0050	2.2408	−0.0357	−0.0112	−0.0245	31.3 (15.5–47.6)
UCLA scale $\times$ CES-D-8	−0.0347	0.3812	−0.0406	−0.0132	−0.0274	32.6 (20.4–49.1)
UCLA scale $\times$ CES-D-7	−0.0347	0.3076	−0.0357	−0.0107	−0.0250	30.0 (18.6–45.7)

Higher scores on the participation measure denote greater participation and higher scores on the outcome denote more depressive symptoms, so the a, c and c' paths are negative; the proportion mediated is a ratio of two negative quantities and is therefore positive.

The coefficient-level result, by contrast, is not sensitive to analytic choices. Across the 336 specifications of Fig. 2 the standardised association ranged from 0.171 to 0.664, and the specifications fall into three bands. The 24 that pair the item with the contaminated outcome had a median of 0.649; the 24 that pair the same item with the de-contaminated outcome had a median of 0.525; the 288 using the UCLA scale had a median of 0.295. The lowest estimate obtained from any doubly contaminated specification (0.591) exceeds the highest estimate obtained from any specification using a clean loneliness measure (0.436), and no combination of covariate set, missing-data rule, participation measure, dichotomisation threshold or weighting brings the two into overlap.

The three bands separate what removing the item from the outcome does from what it does not do. Of the 0.354 difference between the doubly contaminated and the clean medians, 0.124 disappears when the item is taken out of the outcome and is attributable to the algebraic overlap. The remaining 0.230 does not: with a clean outcome the item still yields a higher estimate than every UCLA specification, its lowest value (0.461) exceeding the highest clean value (0.436). That residual gap cannot be attributed to item overlap, and the present design cannot attribute it to anything else. It is consistent with the two instruments differing in ways unrelated to overlap — the item and the CES-D total are answered in the same core interview whereas the UCLA scale is answered later in a self-completed questionnaire, and a single binary item and a three-item scale differ in reliability. Removing the item from the outcome removes the definitional part of the inflation, and only that part.

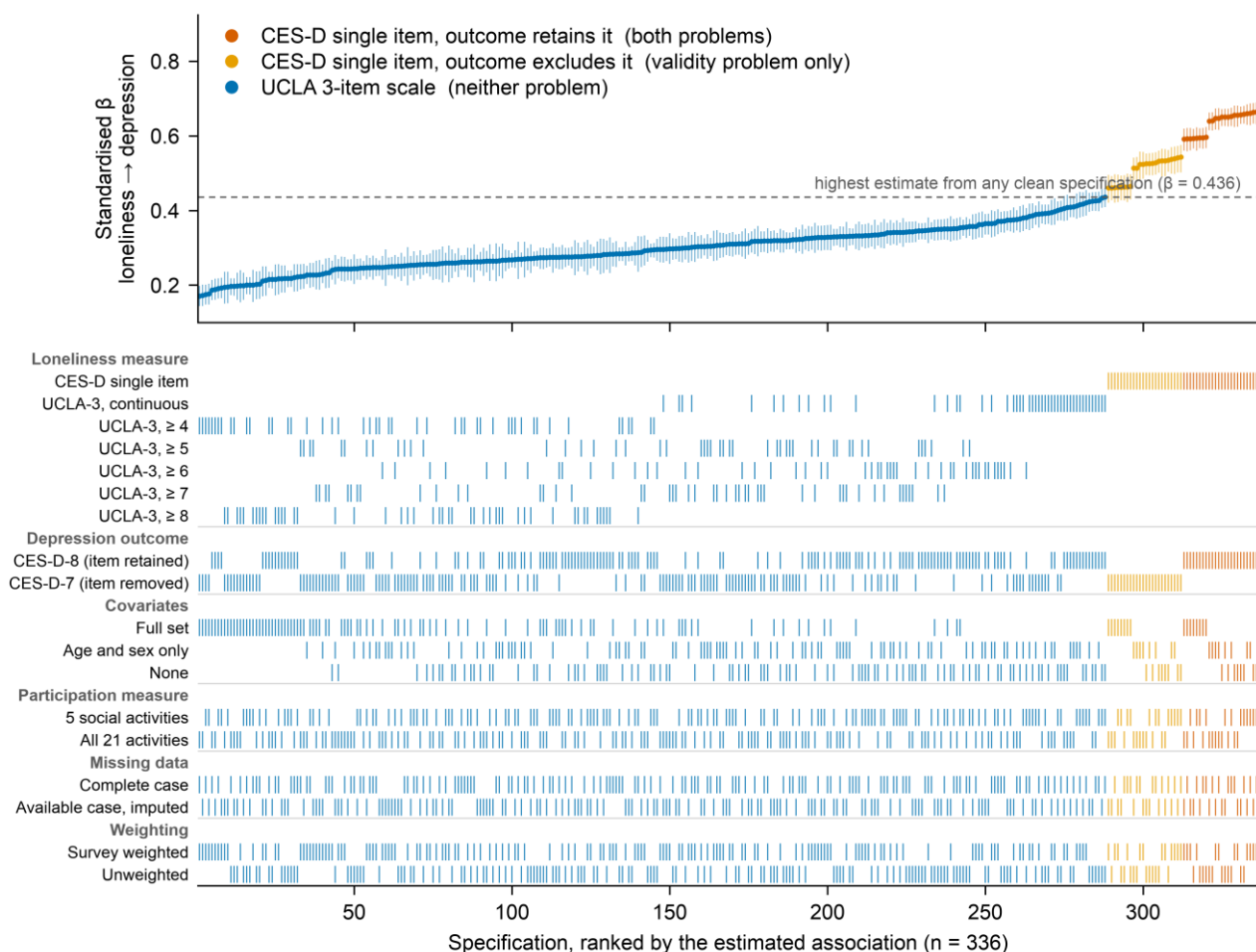


Fig. 2 Specification curve of the loneliness–depression association across 336 combinations of measurement and modelling choices, with Rao–Wu rescaled bootstrap 95% confidence intervals. Estimates are standardised coefficients, ranked in ascending order; the lower panel indicates which choices each specification makes

4. Discussion

4.1. Principal findings and comparison with prior work

The criterion-validity result is a replication. Yu et al. established the low sensitivity of this item in this cohort: validating it against the same three-item reference standard at the same threshold in the 2006 HRS wave, they reported sensitivities of 0.30 and 0.39 by sex against specificities of 0.93 and 0.90, and used those quantities to model how differential underreporting would bias a memory outcome [4]. Twelve years later, in a different wave and a different analytic sample, the corresponding figures are 0.377 and 0.934. The agreement is close. The present analysis adds the statistics that the earlier validation did not report — kappa, predictive values, the area under the curve, and the behaviour of all of these across dichotomisation thresholds — and reports them for their own sake rather than as inputs to a bias model. Because the 2006 and 2018 leave-behind waves were both fielded to subsample A, the respondents partially overlap, and the replication is therefore only partially independent.

A recent comparison in the Canadian Longitudinal Study on Aging reached a different conclusion from a closely similar table. Chamberlain et al. compared the CES-D loneliness item with the same three-item scale in 43,235 respondents and reported that 14% were classified as lonely by both instruments, 67% by neither and 19% discordantly, concluding that the two "provided comparable overall classifications of loneliness", while also finding that respondent characteristics differed across the finer response categories [6]. Two differences in measurement qualify the comparison: their item is drawn from the ten-item CES-D and was answered on a four-level frequency scale, whereas the HRS item is drawn from the eight-item version and is binary. Even so, the present sample produces almost the same distribution under the same tabulation: 9.6% classified lonely by both, 69.7% by neither and 20.7% discordantly, a raw agreement of 79.3% against their 81%. Nearly identical tables therefore support divergent readings depending on which statistic is emphasised. Raw agreement in a population where three-quarters of respondents are not lonely is carried by the concordant negatives; the chance-corrected statistics for the present table are a kappa of 0.363 and a sensitivity of 37.7%. Raw agreement is not an adequate criterion for judging whether a single item can substitute for a loneliness scale in a population in which most respondents are not lonely.

The contamination result connects a long-standing psychometric concern to a measure for which no previous application of it was found. Criterion contamination in depression scales has been discussed for two decades in relation to somatic items in patients with chronic disease, where the same symptoms count towards both the illness and the depression score [8]. The loneliness item raises the identical problem in a different domain, with the additional feature that the contaminating component here is exactly quantifiable. The concern is not confined to a single instrument: loneliness scales share limited variance with one another, with correlations between instruments ranging from 0.384 to 0.777 across 31 countries [13], so the instrument used is a substantive analytic choice rather than a formality.

4.2. Implications

Three practical steps follow, none of which requires new data collection. Where the CES-D loneliness item is used as a loneliness measure, the depression outcome should be the score with that item removed, and results before and after removal should both be reported, so that readers can see the size of the definitional component. Where a single item is validated against a scale, chance-corrected statistics should be reported and raw agreement should not be reported alone. Where a study can afford three additional questions, an independent loneliness scale should be preferred to an item borrowed from the outcome instrument.

4.3. Limitations

The analysis uses a single cross-sectional wave, and the mediation specification in particular cannot support a causal reading; cross-sectional mediation estimates are biased for the longitudinal quantities they are usually taken to represent [14]. The sample represents United States residents aged 50 and

over, and the prevalence figures should not be extrapolated to other populations. The reference standard is a three-item scale rather than a clinical assessment, and the three-item scale is not itself a gold standard, so a discordant classification does not establish that the single item is the one in error; the coefficient decomposition, however, requires no reference standard at all and is unaffected by this limitation.

The two loneliness measures also differ in more than content. The single item is asked during the core interview and the three-item scale in a self-completed questionnaire returned later, so part of the disagreement between them, and part of the residual gap in the specification curve, may reflect the mode and the timing of administration rather than the properties of the item. The leave-behind questionnaire must be completed and returned, and respondents in poorer health are more likely not to return it; the leave-behind weight adjusts for observed differences in response but cannot remove selection on unobserved characteristics. The HRS administers the CES-D in a binary eight-item form, whereas the original scale and several other cohorts use four-level response categories, so the magnitudes reported here should not be transferred to those instruments — the algebraic identity, which depends only on the item being a component of the total, transfers unchanged. The replication of the validity result is only partially independent because the two waves share a subsample. Social participation serves only as an illustrative exposure and its operationalisation varies across studies, which is why alternatives are shown together in the specification curve rather than selected. Finally, the literature search that established the absence of a prior decomposition was conducted at the level of titles and abstracts and did not cover Chinese-language databases, so it supports the statement that no such study was found rather than the stronger claim that none exists.

5. Conclusion

In a nationally representative sample of United States adults aged 50 and over, the CES-D loneliness item fails on both of the requirements that its use presupposes. As a measure of loneliness it identified 37.7% of the respondents whom a validated three-item scale classified as lonely, a result that holds across dichotomisation thresholds and that replicates an earlier validation in the same cohort. As a predictor of the CES-D total score it is not independent of that score: its adjusted coefficient contains a component of exactly one that is present by construction rather than estimated, and that component accounts for 30.9% of the coefficient in the main specification. Studies that measure loneliness with this item and depression with a total score containing it are reporting, in part, an arithmetic identity. Removing the item from the outcome costs nothing and removes that part — though only that part, since the item remains a poor measure of loneliness after the outcome is cleaned.

6. Acknowledgements

The HRS (Health and Retirement Study) is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan. Data collection for the HRS is approved by the University of Michigan Health Sciences and Behavioral Sciences Institutional Review Board (protocol HUM00061128). The present study is a secondary analysis of de-identified public-release data and required no further ethical review. The data products analysed are cited in full as [15]–[18]; the RAND HRS Longitudinal File is cited both in the form required by the Health and Retirement Study [15] and in the form required by the RAND Center for the Study of Aging [16].

References

- [1] Wen Z, Wang H, Liang Q, Liu L, Zhang W, Zhang X. Mediating effect of social support and resilience between loneliness and depression in older adults: A systematic review and meta-analytic structural equation modeling. *Journal of Affective Disorders*, 2024, 365: 246-257.
- [2] Radloff LS. The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. *Applied Psychological Measurement*, 1977, 1(3): 385-401.

- [3] Turvey CL, Wallace RB, Herzog R. A Revised CES-D Measure of Depressive Symptoms and a DSM-Based Measure of Major Depressive Episodes in the Elderly. *International Psychogeriatrics*, 1999, 11(2): 139-148.
- [4] Yu X, Zahodne LB, Gross AL, Needham BL, Langa KM, Cho T, et al. Could differential underreporting of loneliness between men and women bias the gender-specific association between loneliness duration and rate of memory decline? A probabilistic bias analysis of effect modification. *American Journal of Epidemiology*, 2025, 194(3): 811-819.
- [5] Fokkema T, Gierveld JDJ, Dykstra PA. Cross-National Differences in Older Adult Loneliness. *The Journal of Psychology*, 2012, 146(1-2): 201-228.
- [6] Chamberlain SA, Griffith LE, Savage RD, Batara J, Yan W, Gruneir A. Single-item versus scale: Comparing respondent demographic, social, and health characteristics by measure of loneliness using the Canadian Longitudinal Study on Aging (CLSA) data. *PLOS One*, 2026, 21(2): e0341572.
- [7] Sha S, Pan Y, Xu Y, Chen L. Associations between loneliness and frailty among older adults: Evidence from the China Health and Retirement Longitudinal Study. *BMC Geriatrics*, 2022, 22(1): 537.
- [8] Pincus T, Hassett AL, Callahan LF. Criterion Contamination of Depression Scales in Patients with Rheumatoid Arthritis: The Need for Interpretation of Patient Questionnaires (as All Clinical Measures) in the Context of All Information About the Patient. *Rheumatic Disease Clinics of North America*, 2009, 35(4): 861-864.
- [9] Lee SL, Pearce E, Ajnakina O, Johnson S, Lewis G, Mann F, et al. The association between loneliness and depressive symptoms among adults aged 50 years and older: a 12-year population-based cohort study. *The Lancet Psychiatry*, 2021, 8(1): 48-57.
- [10] Salinas J, Beiser AS, Samra JK, O'Donnell A, DeCarli CS, Gonzales MM, et al. Association of Loneliness With 10-Year Dementia Risk and Early Markers of Vulnerability for Neurocognitive Decline. *Neurology*, 2022, 98(13): e1337-e1348.
- [11] Hughes ME, Waite LJ, Hawkey LC, Cacioppo JT. A Short Scale for Measuring Loneliness in Large Surveys: Results From Two Population-Based Studies. *Research on Aging*, 2004, 26(6): 655-672.
- [12] Sonnega A, Faul JD, Ofstedal MB, Langa KM, Phillips JW, Weir DR. Cohort Profile: the Health and Retirement Study (HRS). *International Journal of Epidemiology*, 2014, 43(2): 576-585.
- [13] Newmyer L, Verdery AM, Margolis R, Pessin L. Measuring Older Adult Loneliness Across Countries. *The Journals of Gerontology: Series B*, 2021, 76(7): 1408-1414.
- [14] Maxwell SE, Cole DA. Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods*, 2007, 12(1): 23-44.
- [15] Health and Retirement Study, (RAND HRS Longitudinal File 2022 (V1)) public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor, MI, (2025).
- [16] RAND HRS Longitudinal File 2022 (V1). Produced by the RAND Center for the Study of Aging, with funding from the National Institute on Aging (grant numbers NIA U01AG009740 and NIA R01AG073289) and the Social Security Administration. Santa Monica, CA (May 2025).
- [17] Health and Retirement Study, (2018 HRS Core (Final V2.0)) public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor, MI, (2022).
- [18] Health and Retirement Study, (Cross-Wave Tracker File (Final 2022 V1.0)) public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor, MI, (2025).