Title

Shifting of Cognitive Assessments Between Face-to-Face and Telephone Administration: Measurement Considerations.

Journal

Journal of Gerontology Series B: Psychological Sciences and Social Sciences, 78(2)

Authors

Smith, Jason
Gibbons, Laura
Crane, Paul
et al.

Publication Date

2023-02-19

DOI

10.1093/geronb/gbac135

Peer reviewed

THE GERONTOLOGICAL SOCIETY OF AMERICA®

Journals of Gerontology: Psychological Sciences
cite as: *J Gerontol B Psychol Sci Soc Sci*, 2023, Vol. 78, No. 2, 191–200
https://doi.org/10.1093/geronb/gbac135
Advance Access publication September 13, 2022

OXFORD

Editor's choice

Research Article

Shifting of Cognitive Assessments Between Face-to-Face and Telephone Administration: Measurement Considerations

Jason R. Smith, ScM,[1,*] Laura E. Gibbons, PhD,[2] Paul K. Crane, MD, MPH,[2] Dan M. Mungas, PhD,[3] M. Maria Glymour, ScD, MS,[4] Jennifer J. Manly, PhD,[5] Laura B. Zahodne, PhD,[6] Elizabeth Rose Mayeda, PhD, MPH,[7] Richard N. Jones, ScD,[8] and Alden L. Gross, PhD, MHS[1,9]

[1]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA. [2]General Internal Medicine, University of Washington, Seattle, Washington, USA. [3]Department of Neurology, University of California, Davis, California, USA. [4]Department of Epidemiology and Biostatistics, University of California, San Francisco, California, USA. [5]Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University, New York, New York, USA. [6]Department of Psychology, University of Michigan, Ann Arbor, Michigan, USA. [7]Department of Epidemiology, University of California, Los Angeles Fielding School of Public Health, Los Angeles, California, USA. [8]Department of Psychiatry and Human Behavior, Warren Alpert Medical School, Brown University, Providence, Rhode Island, USA. [9]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA .

*Address correspondence to: Jason R. Smith, ScM, Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St., Baltimore, MD 21202, USA. E-mail: jsmit491@jhu.edu

Received: March 30, 2022; Editorial Decision Date: September 9, 2022

Decision Editor: Alyssa Gamaldo, PhD

Abstract

Objectives: Telephone-administered cognitive assessments are a cost-effective and sometimes necessary alternative to face-to-face assessments. There is limited information in large studies concerning mode effects, or differences in cognition attributable to the assessment method, as a potential measurement threat. We evaluated mode effects on cognitive scores using a population-based sample of community-living older adults.

Methods: We used data from participants aged 65–79 in the 2014 Health and Retirement Study for whom the interview mode was randomized (n = 6,825). We assessed mode differences in test means, whether mode modifies associations of cognition with criterion variables, and formal measurement invariance testing.

Results: Relative to face-to-face assessment, telephone assessment was associated with higher scores for memory and calculation (0.06 to 0.013 standard deviations [*SD*]) and lower scores for nonmemory items (−0.09 to −0.01 *SD*). Cognition was significantly differentially related to instrumental activities of daily living difficulty depending on assessment mode. Measurement invariance testing identified evidence of mode differences in certain tests as a function of mode: adjusting for underlying cognition, the largest mode differences in memory and attention: immediate noun recall, delayed word recall, and serial-7s scores were higher given telephone administration.

Discussion: Differences by mode of administration are apparent in cognitive measurement in older adults, albeit to a small degree in our study, and most pronounced for tests of memory and attention. The importance of accounting for mode differences ultimately depends on one's research question and study sample: not all associations may be affected by mode differences, and such modification may only be apparent among those with lower cognitive functioning.

Keywords: Cognition, Mode effects, Psychometrics, Surveys, Telephone

Cognitive function assessments are commonly used in population-based studies to screen for dementia, track longitudinal changes in cognition, provide a snapshot of cognitive function, or estimate the effects of risk factors on cognitive function or change. Depending on the functional status of the participants or the research question of interest, cognitive function is generally ascertained directly from assessments of the participant (Kirshner et al., 1985).

Most cognitive tests used in large-scale survey research were developed for face-to-face testing (Folstein et al., 1975). Beginning in the late 1980s, telephone adaptations of brief mental status tests were introduced (Brandt et al., 1988). Subsequent studies have adapted telephone-based administration of common cognitive assessments and neuropsychological tests (Bunker et al., 2017; Carlew et al., 2020). Telephone-based assessments are convenient for collecting cognitive function data in large-scale studies, and they also become necessary to continue ongoing studies when face-to-face contact with participants is not safe or feasible, such as during global pandemics or when participants move out of the study catchment area.

It is ideal, of course, in longitudinal studies to keep the mode of assessment constant whenever possible. However, as Robert Burns wrote in 1785, "The best laid plans of mice and men often go awry." Tests are not guaranteed to be equivalent when administered using different modalities (Al Baghal, 2019; Cernat et al., 2016; Ofstedal et al., 2021). While telephone administration is primarily constrained to auditory communication, given difficulties in assessing visual stimuli sent in advance via mail, face-to-face testing leverages both auditory and visual cues. This is potentially a critical distinction, especially in older populations with prevalent sensory impairments. Yet there is limited data comparing face-to-face to telephone-administered cognitive function assessments (Carlew et al., 2020; Castanho et al., 2014; Lachman et al., 2008). Among the existing literature, there is lingering doubt concerning the power of previous studies to detect differences in measurement precision of cognition between mode of test administration, which affects both internal validity (i.e., following adaptation to telephone administration) and comparability across studies (Carlew et al., 2020; Herzog et al., 1997; Lachman et al., 2008).

There are potential pitfalls when comparing cognitive test performance across modes. Tests may be easier or more difficult when administered in one mode or another, generating systematic biases. Or, telephone administration might introduce measurement errors attributable to sensory problems, which could result in differential reliability across the mode. These factors, and others, could manifest in lack of exchangeability of measures across mode or even mode differences in associations of cognition with external variables. To address these pitfalls, in the present study, we examined potential differences in the measurement of cognition administered by telephone or by face-to-face interview. We leveraged cognitive assessment data collected in a population-based sample of community-dwelling older adults aged 65–79 years in the United States in 2014 who were randomly assigned to either face-to-face or telephone modes of assessment. We assessed mode effects on test means and reliability, and evaluated whether mode modifies associations of cognition with other criterion variables.

Method

Data Sources and Study Sample

We used data from the United States Health and Retirement Study (HRS), an ongoing, nationally-representative cohort of adults aged 51 years or older and their spouses in the United States (Sonnega et al., 2014; Staff, 2017). Initiated in 1992, the HRS administers follow-up surveys to all participants biennially and replenishes the cohort approximately every 6 years (i.e., every three waves). The HRS data used in this study is publicly available no Institutional Review Board approval was required.

Our study sample consisted of HRS respondents to the 2014 wave who were aged 65–79 and: (a) were randomized to an interview assignment; and (b) completed cognitive function test items during the 2014 interview (Supplementary Figure 1). We excluded respondents aged less than 65 years because they were not administered all cognitive function tests, and those aged 80 years and older because their interview assignment mode was not randomized (those interviews were always face-to-face). In lieu of direct cognitive assessment, proxy respondents completed a questionnaire concerning change in their respondent's memory over the past 2 years (a modified Informant Questionnaire on Cognitive Decline in the Elderly (Jorm et al., 1988)) for participants too impaired to complete assessments. We excluded these respondents, resulting in a sample size of $N = 6{,}825$.

Survey Mode

In 2014, the HRS administered follow-up surveys via two modes: face-to-face and telephone. The 2014 wave was not one where a new cohort of participants was added to the HRS, thus, this was at least the second assessment wave for most participants (aside from potentially new spouses introduced into the study). The questionnaires were identical for both modes in 2014, with the exception that face-to-face included additional physical measures (e.g., blood pressure, height, and waist measurements) and a leave-behind psychosocial questionnaire. Most face-to-face assessments are conducted in the participant's home. No surveys were completed

over the internet (e.g., video call via camera on a computer) in 2014. Participants could use either land-lines (hard-wired or cordless) or cellular phones (analog or smartphones) to complete telephone surveys.

From 2006 through 2014, continuing participants (including spouses) less than 80 years of age were randomly assigned to either face-to-face or telephone follow-up visits. Respondents could choose, however, to opt out of their random assignment. In the 2014 wave, 7% (n = 499) switched their assignment. For primary analyses of the effects of mode on cognitive test scores, we analyzed all observations "as randomized."

Cognitive Function Assessment

The HRS cognitive function assessment was designed for telephone administration. Thus, it consists solely of verbal measures of cognition. The same cognitive function assessment is administered to all participants, regardless of mode (i.e., telephone or face-to-face). Participants are also given the same instructions for completing the cognitive assessment, irrespective of mode. There are 12 cognitive tests that assess orientation to time (four items), language/naming (four items), memory (two indicators), and attention/concentration (two items). Following HRS convention, we created a summary score that sums points across all tests for a total cognitive score ranging from 0 to 35.

Statistical Analysis

We describe mode effects between telephone and face-to-face administration of cognitive tests by evaluating differences in cognitive test means by mode. Given small imbalances in background characteristics by assessment mode, we present age- and sex-adjusted differences in means (or proportions for binary cognitive test variables), and standardized effect sizes for differences by mode (Cohen's d for continuous variables, Cohen's h for binary variables, and Cohen's w for categorical variables; Cohen, 1988) and significance testing using a two-sided test. In addition to age and sex adjustment, we further adjusted comparisons of means for adherence: not all participants were interviewed in the mode randomly assigned, so the actual difference in probability of a telephone assessment between individuals assigned to telephone versus individuals assigned to face-to-face interviews was only 0.85 (Pr [X = 1|Z = 1] – Pr [X = 1|Z = 0] = 0.932 – 0.078 = 0.854), where X is the actual mode, and Z is the randomized mode. Our primary effect estimates can be considered intent-to-treat estimates of the effect of randomization to telephone on the score. To correct for the nonadherence to random assignment of mode and estimate the effect of completing a telephone interview versus a face-to-face interview, we divided the intent-to-treat estimates by 0.85 (Oakes et al., 2017). This estimates the effect of a telephone interview on a score for people who could have completed either mode, under the likely assumption that random assignment to the telephone only influenced score via the mode of interview.

Because older age and instrumental activity of daily living (IADL) difficulty are associated with poorer cognitive function, as a form of criterion validation, we evaluated whether total cognitive scores were associated with these variables differentially by mode. We regressed age using linear regression (yielding beta coefficients), and count of IADL limitations using negative binomial regression to handle skewness, and parameter estimates are expressed as rate ratios (RR). In both types of models, predictors were total cognitive scores, mode, and an interaction term between mode and total cognitive score. The interaction term tested whether the association of the outcome variable and cognitive functioning differs by mode. The purpose of these regressions was to statistically evaluate the differential relationship between cognitive performance and criterion by mode.

Next, to compare the measurement quality of cognitive functioning by mode, we tested different levels of measurement invariance using a series of nested confirmatory factor analysis (CFA) models of individual cognitive test items. We followed the procedures described by Bontempo and Hofer (2007). Measurement invariance testing can be used to distinguish measurement differences by a mode that are attributable to random error or to systematic error, and to identify specific test items that may be responsible for mode effects.

We evaluated three levels of measurement invariance. Configural invariance is characterized by a two-group CFA of cognitive test items in which all cognitive tests load onto a common factor structure in two groups (face-to-face and telephone), and there are no equality constraints for parameters across the mode. This model serves as a baseline model against which to compare subsequent models that have additional constraints; the model's absolute fit to empirical data is of most interest. The next level for measurement invariance testing is metric invariance, in which the loadings of test items on the cognitive factor are constrained to be equal between modes. Factor loadings can be considered an index of reliability, such that lower values suggest the underlying latent trait is measured with more random error (Bollen, 1989). Thus, metric invariance is akin to evaluating whether there are systematic differences in a random error of test items by mode. The third level of measurement invariance testing is scalar invariance, in which item means (for continuous test items) or thresholds (for categorical test items) are constrained to be equal across groups; this level of invariance is analogous to evaluating differences by mode in terms of systematic error or bias of particular items (Chen, 2008). In addition, we tested for *partial* metric and *partial* scalar invariance by freeing specific item loadings or means/thresholds, respectively. To improve the fit of the single-factor solution, we included a bifactor to account for the correlation between

immediate and delayed word recall; pursuant to invariance testing procedures, loadings for this bifactor were free to vary by mode during configural invariance testing but fixed in subsequent models. In a sensitivity analysis, out of concern that some participants self-selected into a mode, we conducted measurement invariance testing among the subset of participants who adhered to their assigned mode; inferences were largely the same.

After characterizing potential bias in the overall sum score and pinpointing and correcting potentially biased test items via measurement invariance testing, we assessed two ways to correct for any differences by mode in the association of criterion variables, including (a) factor scores estimated from a partially scalar invariant factor, and (b) linear equating by mode in which we used the mean and standard deviation of the total cognitive score among face-to-face administration to equate the telephone administration. In linear equating, cognitive scores in the group administered tests via telephone are adjusted, so that means and standard deviations are constant across modes; details are in Livingston (2014).

We generated descriptive statistics and regressions using Stata version 15.1 (StataCorp, 2017). We conducted measurement invariance testing using Mplus software, version 8.2 (Muthén et al., 1998–2017).

Results

Sample Characteristics

Table 1 shows demographic characteristics, overall and by mode. Participants were on average 72 years old (range 65–79 years). A majority were female (59%) and White (79%), and the average number of years of education was 12.7 years (range 0–17 years). Age did not differ considerably by mode of assessment (72.2 vs 71.9 years); the difference in mean age corresponds to standardized mean difference, *d*, of −0.06. There were no other remarkable demographic differences by mode, with standardized mean differences in the range of 0.00–0.04.

Not all participants adhered to their assigned mode; $n = 275$ participants were assigned a face-to-face interview but did a telephone interview, and $n = 224$ did the opposite. Common reasons for deviation from assigned mode among the $n = 275$ participants interviewed via telephone but who were assigned a face-to-face interview included to persuade a proxy to participate ($n = 185$, 35%); participant hearing problems ($n = 98$, 19%); and interviewer's health, convenience, safety, or weather ($n = 74$, 14%). Common reasons for deviation from assigned mode among the $n = 224$ participants interviewed via face-to-face but who were assigned a telephone interview included participant hearing problems ($n = 77$, 34%); to persuade a proxy to participate ($n = 43$, 19%); and for otherwise unspecified personal reasons ($n = 33$, 15%). People not interviewed in their assigned mode were more likely to be Hispanic (odds ratio [OR] = 1.71, 95% confidence interval [CI] = 1.33, 2.21), and have less education (OR = 0.96 per year of school, 95% CI = 0.93, 0.98).

Cognitive Function Assessment

Table 2 shows the age- and sex-adjusted test performance for the total cognitive score and individual items by mode. Ranges of scores did not differ by mode. We did not observe large differences in the total cognitive summary score or among individual test items. Tests of episodic memory had on average higher means from telephone evaluation (e.g., immediate noun recall: $d = 0.05$, delayed noun recall: $d = 0.11$) while some nonmemory items had slightly higher means during face-to-face administration (orientation to day of week: $h = -0.07$; vice president's name: $h = -0.08$), however all effect sizes were less than 0.2, including for the total cognitive score ($d = 0.05$). These differences did not change appreciably when additionally adjusted for adherence to the assigned mode.

Association Between Cognitive Performance and Mode Assignment on Age and IADL Variables

We tested for differences in the association of total cognitive score by mode with each criterion variable, age, and IADL difficulty, separately. As expected, lower cognitive performance was associated with both older age (β = −0.79; 95% CI = −1.11, −0.47) and greater IADL difficulty (RR = 0.60; 95% CI = 0.48, 0.74). While we did not detect an interaction between cognition and mode for age (β = 0.02; 95% CI = −0.19, 0.22), there was an interactive effect of mode with cognition in predicting the IADL difficulty sum score (RR = 0.87; 95% CI = 0.76, 0.99; Supplementary Table 1). These results indicate higher levels of cognition were associated with a lower risk of IADL difficulty among both participants who completed face-to-face interviews or telephone interviews, but that the association was stronger in the latter group. Subsequent interrogation of this interaction revealed the mode disparity is largest at lower levels of cognition (e.g., among those with a total cognition score of 7 and lower, representing 1.4% of the sample; Figure 1, panel A).

Detecting Measurement Differences by Mode Using Measurement Invariance

Next, we assessed configural invariance and tested metric and scalar levels of measurement invariance (reported in Supplementary Table 2). Configural invariance is confirmed if a multiple group CFA model without cross-group parameter constraints and the same factor loading pattern (which items load in which factors) fits adequately. This assessment is not a formal hypothesis test, and the principal role of the configural model is to serve as a baseline model to compare more restrictive invariance testing models (Bontempo & Hofer, 2007). The fit of the one-factor configural model

Table 1. Characteristics of HRS Participants Aged 65–79 Years in 2014 Assigned to Face-to-Face and Telephone Cognitive Function Assessment Administration (*N* = 6,825)

Characteristic	Overall (*N* = 6,825)	Face-to-face assignment (*n* = 3,528)	Telephone assignment (*n* = 3,297)	Effect size for difference in means or proportions by assignment[a]
Age, mean (*SD*; range), *y*	72.0 (4.3; 65–79)	71.9 (4.3)	72.2 (4.3)	–0.06
Female sex, *n* (%)	4019 (59)	2040 (58)	1979 (60)	–0.04
Participant race, *n* (%)[b]				0.00
White	5,368 (79)	2,775 (79)	2,593 (79)	
Black	1,098 (16)	570 (16)	528 (16)	
All other racial groups	354 (5)	182 (5)	172 (5)	
Not Hispanic or Latino, *n* (%)[c]	6,108 (90)	3,160 (90)	2,948 (90)	0.00
Number of years in school, mean (*SD*; range)	12.7 (3.1; 0–17)	12.7 (3.1)	12.7 (3.1)	0.00
Sum of IADLs unable to complete, *n* (%)[d]				0.00
0	5,908 (87)	3,036 (86)	2,872 (87)	
1	522 (8)	295 (8)	227 (7)	
2	219 (3)	107 (3)	112 (3)	
3	92 (1)	50 (1)	42 (1)	
4	51 (1)	24 (1)	27 (1)	
5	33 (1)	16 (1)	17 (1)	

Notes: Characteristics of the study sample. HRS = Health and Retirement Study; IADL = instrumental activity of daily living; *SD* = standard deviation.
[a]Effect size statistics (Cohen's *d* for continuous variables, *h* for binary variables, and *w* for categorical variables) for the difference between face-to-face and telephone cognitive assessment assignment.
[b]Missing data of race among face-to-face (*n* = 1) and telephone assignment (*n* = 4) due to nonreporting among participants.
[c]Comparing non-Hispanic/Latino to Hispanic/Latino; missing data of ethnicity among face-to-face (*n* = 1) and telephone assignment (*n* = 2) due to nonreporting among participants.
[d]IADLs participants are unable to complete due to a health or memory problem.

was excellent (root mean squared error of approximation, RMSEA = 0.022; confirmatory fit index, CFI = 0.996; standardized root mean squared residual, SRMR = 0.059). Fit of the metric model, in which loadings were fixed to be the same by mode, was also excellent (RMSEA = 0.020; CFI = 0.997; SRMR = 0.060) and, importantly, was not significantly worse than the configural model ($\Delta\chi^2 = 12.37$, df = 8, $p = .14$).

The fit of the scalar model, in which thresholds were fixed equal across modes, was excellent (RMSEA = 0.021; CFI = .995; SRMR = 0.060) although it fit significantly worse than the metric model ($\Delta\chi^2 = 78.51$, df = 26, $p < .001$) suggesting systematic differences in at least some cognitive test thresholds by mode. Therefore we examined which thresholds differed most by mode (based on the metric invariance model), and estimated a partial scalar-invariant model in which thresholds for immediate and delayed word recall, serial 7s subtraction, and knowing the day of the week were allowed to vary by mode. Fit relative to the metric model was comparable ($\Delta\chi^2 = 8.78$, df = 4, $p = .07$), suggesting that the difficulty of the four selected cognitive test items differed by mode of administration.

The effect of item-level differences identified through tests of measurement invariance is encapsulated in Table 3. Table 3 shows differences in the mean level of cognitive functioning by mode, assuming configural invariance, metric invariance, scalar invariance (which makes the same assumptions regarding equivalence of items by mode as summing items does), and a partial scalar-invariant model (which allows for differences in the difficulty of the four items by mode, but otherwise enables appropriate comparison across mode). Means are standardized to a N(0,1) distribution in the face-to-face mode, allowing us to interpret differences on an effect size scale relative to the face-to-face mode's standard deviation. The mean level of cognitive functioning assuming a scalar invariance model was $d = 0.12$ standard deviation units higher for telephone administration compared to face-to-face ($p < .001$); this absolute difference is attenuated by 63% to $d = -0.09$ standard deviation units in the partial scalar-invariant model ($p = .12$). This result means that differences in score by mode—which participants were randomized to—are reduced after allowing certain cognitive test thresholds (immediate and delayed word recall, serial 7s subtraction, and knowing the day of the week) to vary by mode.

Potential Strategies to Mitigate Detected Measurement Differences by Mode in Criterion Variables

Finally, we evaluated the extent to which potential solutions to measurement differences by mode mitigated the interaction of mode with cognition in the relationship with IADL difficulty (Table 4). Correction of the total cognitive

Table 2. Mean Cognitive Function Assessment Item Scores Among HRS Participants Aged 65–79 Years in 2014 Between Face-to-face and Telephone Administration (N = 6,825)

Test item	Domain	Face-to-face assignment mean (*SE*)	Telephone assignment mean (*SE*)	Age and sex adjusted difference[a] (*SE*)	Age and sex adjusted effect size	Age, sex, and adherence adjusted effect size
"What month are we in?"[b]	Orientation	0.98 (0.18)	0.98 (0.18)	−0.00 (0.15)	−0.02	−0.02
"What is the year we are in?"[b]	Orientation	0.99 (0.21)	0.99 (0.21)	−0.00 (0.16)	−0.03	−0.04
"What is today's date?"[b]	Orientation	0.87 (0.07)	0.88 (0.08)	0.01 (0.07)	0.03	0.04
"What day of the week is today?"[b]	Orientation	0.99 (0.18)	0.98 (0.17)	−0.01 (0.15)	−0.07	−0.08
"What do people usually use to cut paper?"[b]	Language/naming	0.99 (0.30)	0.99 (0.30)	−0.00 (0.25)	−0.00	−0.01
"What do you call the prickly green plant that lives in the desert?"[b]	Language/naming	0.99 (0.19)	0.98 (0.19)	−0.00 (0.16)	−0.02	−0.02
Who is the President of the United States right now?"[b]	Language/naming	1.00 (0.45)	1.00 (0.41)	−0.00 (0.36)	−0.04	−0.05
"Who is the Vice President?"[b]	Language/naming	0.93 (0.11)	0.90 (0.11)	−0.02 (0.09)	−0.08	−0.09
"Count backwards from 20 to 1"[c]	Attention/ concentration	0.94 (0.11)	0.93 (0.11)	−0.01 (0.10)	−0.03	−0.04
Serial 7s[d]	Attention/ concentration	3.77 (0.04)	3.85 (0.04)	0.08 (0.04)	0.05	0.07
Immediate noun recall[e]	Memory	5.21 (0.03)	5.28 (0.03)	0.07 (0.03)	0.05	0.06
Delayed noun recall[e]	Memory	4.45 (0.04)	4.64 (0.04)	0.19 (0.04)	0.11	0.13
Total cognitive score	Total	21.62 (0.12)	21.85 (0.12)	0.23 (0.11)	0.05	0.06

Notes: Effect sizes are Cohen's h for test items scored from 0 to 1 (binary), and Cohen's d effect size calculated for test items scored from 0–2, 0–5, or 0–10 (continuous). Covariate adjustment was done by regressing the cognitive score on assigned mode and covariates. HRS = Health and Retirement Study; *SE* = standard error.

[a]Difference = telephone assignment mean score – face-to-face assignment mean score.

[b]Scored from 0 to 1.

[c]Scored from 0 to 2.

[d]Scored from 0 to 5.

[e]Scored from 0 to 10.

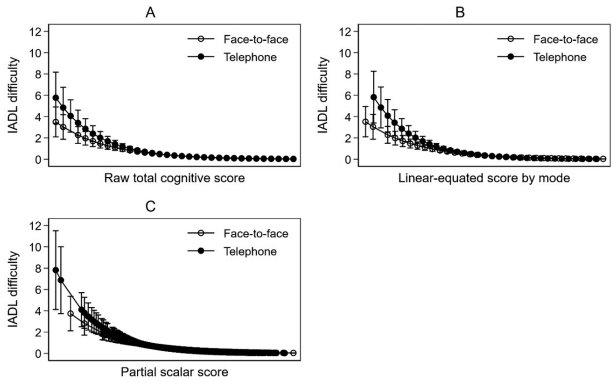

Figure 1. Interaction between mode and several alternatives for calculating cognitive performance in predicting IADL difficulty (N = 6,825). Each panel shows the model-estimated association between IADL difficulty (count of difficulties) and cognition, where cognition is defined by the raw sum score of all items (panel A), a linear-equated score by mode (panel B), and a factor score from a partially scalar-invariant confirmatory factor analysis (panel C). IADL = instrumental activities of daily living.

score via linear equating by mode, which produces a score with the same mean and standard deviation, by design does not eliminate the significant interaction effect (Table 4 and Figure 1, panel B). Use of a factor score, from a partially scalar-invariant model to correct the total cognitive score, produced an identical point estimate for the interaction effect, although the interaction was no longer statistically significant (RR = 0.87, 95% CI = 0.75, 1.01; Table 4; Figure 1, panel C). These results suggest none of the approaches we tried were able to entirely remove the bias introduced by the mode of administration.

Discussion

In this study, we evaluated mode effects in cognitive item scores, and overall cognitive function score, among a nationally representative sample of older adults aged 65–74 randomized to either face-to-face or telephone assessments. Findings indicate differences in mean scores by mode: higher average memory test scores among telephone assessments, and higher nonmemory average test scores among face-to-face assessments. These differences were not appreciably larger than age differences by mode in the sample. However, for many tests, the mode differences persisted when adjusting for age, sex, and adherence to the assigned mode. We did find that mode differences modified associations between cognitive performance and an important criterion variable, IADL difficulty. Measurement invariance testing revealed that generally, among the cognitive tests administered, tests of memory and attention were most susceptible to mode differences.

The importance of mode effects will depend on one's research question and study sample. On the one hand, the mode effects are very small and should be interpreted within the context of the study sample. However, even

Table 3. Estimated Mode Differences in Scalar-Invariance and Partial Scalar-Invariant Factors (N = 6,825)

Model	Difference in latent factor mean (*SE*)[a]	*p* Value
Configural	0.18 (0.13)	.10
Metric	0.07 (0.08)	.43
Scalar[b]	0.12 (0.04)	<.001
Partial scalar[c]	−0.09 (0.06)	.12

Note: *SE* = standard error.
[a]Difference = telephone administration mean − face-to-face administration mean.
[b]Item loadings or thresholds are constrained by mode.
[c]Item thresholds are allowed to vary by mode.

subtle effects attributable to mode can be comparable in magnitude to effects of risk factors important to population health. Thus, mode effects should be recognized and modeled when analyzing data collected from a mixture of modes. Importantly, in our sample, mode differences are greatest among respondents with lower levels of cognitive functioning (Figure 1). Participants with low total cognitive scores (7 points and lower out of a possible total score of 35) represented a small number of individuals in our sample. However, in other samples or for scientific questions that leverage more selected populations, such as clinical samples or samples of cognitively impaired older adults, mode differences might overwhelm substantively important signals. Thus, investigators should carefully consider their study sample and scientific question before switching modes of data collection.

Assuming that one's research question requires careful consideration to mode differences if they are present, a question arises as to how cognitive data from different modes should be treated in models. In our study, we evaluated several potential solutions, the simplest being to include a covariate for mode. We also applied linear equating, which has been used to correct for parallel but nonequivalent alternate forms of memory tests (Briceño et al., 2021; Gross et al., 2012, 2019), which presents a similar problem. None of these approaches entirely removed the bias. If one is truly worried about mode effects in their data, one potential suboptimal resolution, not evaluated here, might be to collapse continuous cognitive scores into categorical scores for cognitive impairment to mask measurement bias by mode; such a coarsening of course loses a great deal of information. Thus, further research is needed to better understand how to adjust for different modes of presentation; this solution may very well be unique to certain study designs.

The cognitive tests most susceptible to mode differences were immediate and delayed noun recall, and serial 7s: in all cases, scores were higher (indicating they were easier) when administered via telephone. For these tests, in particular, participants might have leveraged strategies not intended by the investigator, such as writing down words (instead of remembering) or writing the calculations

Table 4. Negative Binomial Regression of IADL Sum on Total Cognitive Function Score (N = 6,825)

	IADL difficulty	
Covariate	RR	95% CI
Total cognitive score (raw)		
Cognition	0.60	(0.48, 0.74)
Telephone mode (vs face-to-face)	0.87	(0.74, 1.01)
Interaction	0.87	(0.76, 0.99)
Total cognitive score (linear-equated by mode)		
Cognition	0.60	(0.48, 0.74)
Telephone mode (vs face-to-face)	0.84	(0.72, 0.98)
Interaction	0.83	(0.72, 0.95)
Partial scalar score		
Cognition	0.59	(0.47, 0.74)
Telephone mode (vs face-to-face)	0.77	(0.66, 0.90)
Interaction	0.87	(0.75, 1.01)

Notes: This table presents possible approaches to correct for differences by mode in the association of the criterion variable of IADL difficulty. In the first approach, mode is included as a covariate, and the primary covariate is a raw sum of cognitive test items. In the second approach, the total cognitive score outcome is adjusted via linear equating by mode, which produces a score with the same mean and standard deviation. In the third approach, the total cognitive score primary covariate is a factor score from a partially scalar-invariant model that corrects for mode effects in cognition. For each approach, the coefficient for the primary covariate is provided alongside the mode term, and their interaction. As shown by the interactions, all of these approaches fail to account for mode differences. IADL = instrumental activities of daily living; RR = rate ratio; CI = confidence interval.

(instead of mentally calculating). This should not necessarily be termed cheating, because in the HRS the directions to the tasks over the telephone did not explicitly ask participants not to write down answers to questions or use external aids, but if even a fraction of the participants did adopt such strategies, that might explain higher average scores. Previous studies, however, suggest that overt cheating on tests is rare (Lachman et al., 2008); this would also likely result in a greater difference in scores than what we observed, so this is not the most likely explanation. In general, several factors could potentially influence performance on cognitive tests, for example, the time of day of the assessment, difficulty hearing verbal instructions, or distractions. Of course, performance on challenging tasks might be easier on a telephone because the participant is in a familiar context, or because it removes other potentially stressful conditions associated with face-to-face interviews (Sindi et al., 2013), such as having a stranger perform the interview in the participant's home. In the case of anxiety, distraction, or general apprehension and confusion among those not accustomed to testing, it is advisable to reach out to participants beforehand to review strategies to minimize distractions, and to remind them that no one is expected to get a perfect score (Lachman et al., 2008).

Our findings are largely consistent with existing literature. In a sample of 110 women aged 65–90 randomly assigned to face-to-face or telephone administration of the Telephone Interview for Cognitive Status-Modified, Rapp et al. (2012) found similar differences in mean verbal learning scores comparing telephone to face-to-face administration (e.g., long delay free-recall: d = –0.06). The effect size for the difference in mean global cognitive functioning score comparing telephone to face-to-face was d = –2.55 in Rapp et al.'s study, indicating higher mean scores for face-to-face administration. That study's mode difference was considerably higher than what we found in the HRS; the prior study used different cognitive measures and had a smaller sample size of only women who were tested in an academic medical setting, which may account for differences in magnitudes of mode effects between studies. That study also reported an effect of mode on rates of change over 6 months for the long delay free-recall measure, which our study did not evaluate.

Although we find evidence of differences in overall cognitive performance under different assessment modes, an important caveat is that our findings ought to be replicated in studies using more expansive neuropsychological performance tasks because the findings may not hold for batteries designed for face-to-face assessment and adapted for telephone administration. Given extrinsic forces such as global pandemics, other epidemiologic studies may be forced to shift from face-to-face to remote assessment, or even replace existing tests with measures designed for remote assessment (this could also be the case when copyrights make adaptation impermissible). In the event that data collection protocols must change, diligence is warranted with respect to the selection of cognitive tests, clear documentation of which mode is used for testing, and the threats to validity that flow from protocol modification.

Moreover, in studies for which face-to-face tests were adapted for telephone administration as a protocol change, continuing remote assessment may be advisable even after face-to-face assessment is once again feasible in order to create high-quality calibration samples. For example, when it is safe to return to face-to-face assessment, investigators might consider randomizing participants to return to face-to-face versus continuing on telephone administration. While costly and administratively burdensome, such study designs may be necessary to obtain data on persons administered cognitive function tests in both face-to-face and remote modes, in random and counterbalanced order, in order to deal with some sources of bias (e.g., retest effects).

We acknowledge several limitations of our study. Our sample consisted of predominately cognitively unimpaired respondents, with few respondents in the cognitively impaired or lower levels of performance. This select composition could potentially hinder inferences made not only to participants starting out as cognitively impaired, but also to those that transition from unimpaired to impaired, or to older age groups (i.e., 80+ years) not included in our study. Another limitation is that there are inherent drawbacks to

any measure of global cognition obtained from telephone administration. Tests of psychomotor performance, visualization, and executive functions such as speeded tasks are not readily captured over a telephone. Face-to-face interviews, and to some extent, video-based testing (Marra et al., 2020), provide a more fertile environment to measure both nonverbal cognitive measures and motor skills. An additional limitation is that the primary criterion variable we used to compare associations of cognition across modes was IADL difficulty. Self-reported IADL questions also varied by mode, however.

A further study limitation is that there are specific methodological tradeoffs in the HRS. The HRS did not adopt a rigorous within-person study design with counterbalancing of the order of mode, and randomization yielded some small demographic differences between groups that could have affected inferences. And the cognitive tests in the HRS—a subset of the Telephone Interview for Cognitive Status (Brandt et al., 1988)—were selected for brevity and to be administered in conjunction with several other surveys by lay interviewers. Many components of cognition are condensed into a few cognitive tests. This smaller number of items could potentially decrease measurement precision of some dimensions and narrow the range of observable data, particularly when compared to other cognitive batteries designed for telephone administration (Tun et al., 2006). Prior analysis of the measurement properties of the HRS-specific tests, however, have demonstrated reasonable dimensionality and internal consistency for an overall cognitive functioning score (used for the configural model in the present study), a memory factor, and a mental status factor (McArdle et al., 2007). This suggests reasonable internal validity; and for the purposes of longitudinal studies interested in overall cognitive function, the HRS cognitive interview is a pragmatic option. These tests are descriptive in terms of differentiating between cognitively normal and impaired, providing reasonable measures for overall cognitive functioning and, critically, memory.

Conclusions

Leveraging a large, nationally representative sample of community-dwelling U.S. older adults in 2014 who were randomized to the mode of assessment of cognitive function, we found mode effects on cognitive test score means. Among those with lower cognitive function, these differences by mode were large enough to modify associations between cognitive function and IADL difficulty. Accounting for measurement differences introduced by mode will ultimately depend on the research question and study sample of an investigation that uses mixed modes of assessments.

Supplementary Material

Supplementary data are available at *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences* online.

Funding

This work was supported by the National Institutes of Health (R01-AG051170, P50 AG05136 [L. E. Gibbons], and R00AG053410 [E. R. Mayeda]). The Health and Retirement Study is sponsored by the National Institutes of Health (U01AG009740) and is conducted by the University of Michigan.

Conflict of Interest

None declared.

Acknowledgments

We thank the staff and participants of the Health and Retirement Study for their important contributions. This study was not preregistered. This study was based on publicly available data from the Health and Retirement Study dataset (data available at: https://hrs.isr.umich.edu/). Portions of this work were presented to the MELODEM working group on April 16, 2020 and benefited from the helpful discussion that followed. Advanced Psychometrics in Cognitive Aging Working Group; P. K. Crane (University of Washington, pcrane@uw.edu); L. E. Gibbons (University of Washington, gibbonsl@uw.edu); A. L. Gross (Johns Hopkins University, agross14@jhu.edu); M. M. Glymour (UCSF, mglymour@gmail.com); R. N. Jones (Brown University, Rich_Jones@brown.edu); J. J. Manly (Columbia University, jjm71@cumc.columbia.edu); E. R. Mayeda (UCLA, ermayeda@ph.ucla.edu); D. M. Mungas (UC Davis, dmmungas@ucdavis.edu); L. B. Zahodne (University of Michigan, lzahodne@umich.edu).

References

Al Baghal, T. (2019). The effect of online and mixed-mode measurement of cognitive ability. *Social Science Computer Review*, **37**(1), 89–103. doi:10.1177/0894439317746328

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods and Research*, **17**(3), 303–316. doi:10.1177/0049124189017003004

Bontempo, D. E., & Hofer, S. M. (2007). Assessing factorial invariance in cross-sectional and longitudinal studies. In A. D. Ong & M. H. M. van Dulmen (Eds.), *Oxford handbook of methods in positive psychology* (pp. 153–175). Oxford University Press.

Brandt, J., Spencer, M., & Folstein, M. (1988). The telephone interview for cognitive status. *Neuropsychiatry, Neuropsychology and Behavioral Neurology*, **1**(2), 111–117. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=psyh&AN=1990-29825-001&site=ehost-live&scope=site

Briceño, E. M., Gross, A. L., Giordani, B. J., Manly, J. J., Gottesman, R. F., Elkind, M. S. V., & Levine, D. A. (2021). Pre-statistical considerations for harmonization of cognitive instruments: Harmonization of ARIC, CARDIA, CHS, FHS, MESA, and NOMAS. *Journal of Alzheimer's Disease*, **83**(4), 1803–1813. doi:10.3233/jad-210459

Bunker, L., Hshieh, T. T., Wong, B., Schmitt, E. M., Travison, T., Yee, J., & Inouye, S. K. (2017). The SAGES telephone neuropsychological battery: Correlation with in-person measures. *International Journal of Geriatric Psychiatry*, **32**(9), 991–999. doi:10.1002/gps.4558

Carlew, A. R., Fatima, H., Livingstone, J. R., Reese, C., Lacritz, L., Pendergrass, C., Bailey, K. C., Presley, C., Mokhtari, B., & Cullum, C. M. (2020). Cognitive assessment via telephone: A scoping review of instruments. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, **35**(8), 1215–1233. doi:10.1093/arclin/acaa096

Castanho, T. C., Amorim, L., Zihl, J., Palha, J. A., Sousa, N., & Santos, N. C. (2014). Telephone-based screening tools for mild cognitive impairment and dementia in aging studies: A review of validated instruments. *Frontiers in Aging Neuroscience*, **6**, 16. doi:10.3389/fnagi.2014.00016

Cernat, A., Couper, M. P., & Ofstedal, M. B. (2016). Estimation of mode effects in the Health and Retirement Study using measurement models. *Journal of Survey Statistics and Methodology*, **4**(4), 501–524. doi:10.1093/jssam/smw021

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, **95**(5), 1005–1018. doi:10.1037/a0013193

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. doi:10.4324/9780203771587

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, **12**(3), 189–198. doi:10.1016/0022-3956(75)90026-6

Gross, A. L., Inouye, S. K., Rebok, G. W., Brandt, J., Crane, P. K., Parisi, J. M., Tommet, D., Bandeen-Roche, K., Carlson, M. C., & Jones, R. N. (2012). Parallel but not equivalent: Challenges and solutions for repeated assessment of cognition over time. *Journal of Clinical and Experimental Neuropsychology*, **34**(7), 758–772. doi:10.1080/13803395.2012.681628

Gross, A. L., Kueider-Paisley, A. M., Sullivan, C., & Schretlen, D. (2019). Comparison of approaches for equating different versions of the mini-mental state examination administered in 22 studies. *American Journal of Epidemiology*, **188**(12), 2202–2212. doi:10.1093/aje/kwz228

Herzog, A. R., & Wallace, R. B. (1997). Measures of cognitive functioning in the AHEAD Study. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, **52**, 37–48. doi:10.1093/geronb/52b.special_issue.37

Jorm, A. F., & Korten, A. E. (1988). Assessment of cognitive decline in the elderly by informant interview. *British Journal of Psychiatry*, **152**, 209–213. doi:10.1192/bjp.152.2.209

Kirshner, B., & Guyatt, G. (1985). A methodological framework for assessing health indices. *J Chronic Dis*, **38**(1), 27–36. doi:10.1016/0021-9681(85)90005-0

Lachman, M. E., & Tun, P. A. (2008). Cognitive testing in large-scale surveys: Assessment by telephone. In S. M. A. D. F. Hofer (Ed.), *Handbook of cognitive aging: interdisciplinary perspectives* (pp. 506–523). SAGE.

Livingston, R. A. (2014). *Equating test scores (Without IRT)*. Available at: https://www.ets.org/Media/Research/pdf/LIVINGSTON2ed.pdf. Accessed 1 December, 2021.

Marra, D. E., Hamlet, K. M., Bauer, R. M., & Bowers, D. (2020). Validity of teleneuropsychology for older adults in response to COVID-19: A systematic and critical review. *Clinical Neuropsychologist*, **34**(7–8), 1411–1452. doi:10.1080/13854046.2020.1769192

McArdle, J. J., Fisher, G. G., & Kadlec, K. M. (2007). Latent variable analyses of age trends of cognition in the Health and Retirement Study, 1992–2004. *Psychology and Aging*, **22**(3), 525–545. doi:10.1037/0882-7974.22.3.525

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Oakes, J. M., & Kaufman, J. S. (2017). *Methods in social epidemiology*. John Wiley & Sons, Incorporated.

Ofstedal, M. B., McClain, C. A., & Couper, M. P. (2012). Measuring cognition in a multi-mode context. In P. Lynn (Ed.), *Advances in longitudinal survey methodology*. Wiley. doi:10.1002/9781119376965.ch11

Rapp, S. R., Legault, C., Espeland, M. A., Resnick, S. M., Hogan, P. E., Coker, L. H., & Shumaker, S. A. (2012). Validation of a cognitive assessment battery administered over the telephone. *Journal of the American Geriatrics Society*, **60**(9), 1616–1623. doi:10.1111/j.1532-5415.2012.04111.x

Sindi, S., Fiocco, A. J., Juster, R. P., Pruessner, J., & Lupien, S. J. (2013). When we test, do we stress? Impact of the testing environment on cortisol secretion and memory performance in older adults. *Psychoneuroendocrinology*, **38**(8), 1388–1396. doi:10.1016/j.psyneuen.2012.12.004

Sonnega, A., Faul, J. D., Ofstedal, M. B., Langa, K. M., Phillips, J. W., & Weir, D. R. (2014). Cohort profile: The Health and Retirement Study (HRS). *International Journal of Epidemiology*, **43**(2), 576–585. doi:10.1093/ije/dyu067

Staff, H. R. S. (2017). *Sample sizes and response rates*. Retrieved from https://hrs.isr.umich.edu/sites/default/files/biblio/ResponseRates_2017.pdf

StataCorp. (2017). *Stata Statistical Software: Release 15*. StataCorp LLC.

Tun, P. A., & Lachman, M. E. (2006). Telephone assessment of cognitive function in adulthood: The Brief Test of Adult Cognition by Telephone. *Age and Ageing*, **35**(6), 629–632. doi:10.1093/ageing/afl095