HRS | HEALTH AND RETIREMENT STUDY
A Longitudinal Study of Health, Retirement, and Aging
Sponsored by the National Institute on Aging

Measuring Cognition in a Multi-mode Context

Documentation Report

Colleen A. McClain,
Mary Beth Ofstedal
Mick P. Couper

Survey Research Center
Institute for Social Research
University of Michigan
Ann Arbor, Michigan

May 2018

Funding

The Health and Retirement Study is funded by a grant from the National Institute on Aging (U01 AG009740) with supplemental support from the Social Security Administration. HRS is conducted by the University of Michigan.

Suggested Citation

McClain, C. A., Ofstedal, M., B& Couper, M. P. (2018). Measuring Cognition in a Multimode Context. University of Michigan.
https://hrs.isr.umich.edu/publications/biblio/9606

Acknowledgements

Data used in the study come from the Health and Retirement Study, which is funded by a grant from the National Institute on Aging (U01 AG009740) with supplemental support from the Social Security Administration. HRS is conducted by the University of Michigan.

INTRODUCTION

As large-scale surveys that have traditionally been administered via telephone or face-to-face modes increasingly move toward including a web option, challenges arise in adapting to self-administration. Striking a balance between taking advantage of the opportunities of the self-administered, computer-based mode and maximizing comparability with interviewer-administered modes presents operational and substantive challenges for survey researchers. This is a particular concern in the context of a longitudinal study when the introduction of a new mode may disrupt time series estimates of trends and trajectories that are of primary value in longitudinal studies.

Although multi-mode studies may present challenges for a variety of survey measures, tests of cognitive ability are especially challenging. Measures of cognitive ability have been incorporated in many population-based surveys, and they are especially common in longitudinal studies of health and aging. These measures can be methodologically challenging to administer even without the complication of mixing modes (Herzog et al., 1999), and the introduction of a new mode adds further complexity. In some cases, the tests that have formed the core of interviewer-administered research designs are difficult or impossible to administer in an online setting, raising questions about how to design measures that minimize measurement error and respondent difficulty while maximizing comparability and response quality across modes. Furthermore, and of particular relevance for studies of aging across the world, these issues may be exacerbated for older respondents who may be unfamiliar with technology or have cognitive impairments that could affect the quality and completeness of the data differentially across modes. Despite these challenges and the shift of many longitudinal studies to web administration, there are few mode comparisons that focus on cognitive measures. Thus, the implications of mixed-mode design decisions for the measurement of cognition in longitudinal surveys still remain largely unclear.

Using data from the Health and Retirement Study (HRS), collected from the same respondents over the course of three years via web, telephone (CATI), and face-to-face (CAPI) administration, we address the following questions: What are the implications of mixing modes for measurement of cognitive performance in a longitudinal setting? Do the same tests administered to the same individuals in different modes produce different response distributions

and response behavior (for example, differences in selection of non-substantive answers or response times)? Finally, are any of the observed mode differences consequential for the substantive conclusions that would be drawn?

The analysis focuses in particular on the implications of transitioning from interviewer-based administration to web administration during the course of an ongoing panel study. Comparisons of measurement error in interviewer-administered and web modes have become more common in recent years (Dillman & Christian, 2005; Duffy, Smith, Terhanian, & Bremer, 2005; Fricker, Galesic, Tourangeau, & Yan, 2005; Chang & Krosnick, 2009; Heerwegh, 2009; Cernat, Couper, & Ofstedal, 2016); however, few studies have included tests of cognitive ability in their comparisons (for a recent exception, see Al Baghal, 2017). We conduct an initial examination of several cognitive tests, with the goals of outlining considerations for test selection in future mixed-mode studies and highlighting particular areas of concern when mixing modes in studies that measure cognitive ability or decline as a key outcome or predictor.

We first review the existing uses of cognitive measures in large-scale surveys; discuss the contribution of mixed-mode research in survey methodology generally; and highlight particular sources of insight for mode comparisons involving cognitive assessments. We then describe the data used and analyses undertaken. Finally, we present results and discuss their implications and future directions for research, highlighting places where more work is needed.

MOTIVATION AND PREVIOUS LITERATURE

Measurement of Cognition in Surveys

As research on the role of cognitive ability and change in predicting a variety of health, economic, and social outcomes becomes increasingly common, large-scale surveys have moved to include or expand cognitive assessments in their instruments. Examples of tests utilized in other survey contexts can be drawn from existing household surveys, as well as from governmental and private resources that specialize in self-administered tests for touchscreen and online implementation.

Cognitive assessments are an integral part of a number of large-scale longitudinal surveys. For example, the HRS and its sister studies, such as the Survey of Health, Ageing and Retirement in Europe (SHARE) and the English Longitudinal Study of Ageing (ELSA) among others, contain extensive batteries of cognitive tests. Cognition and Aging in the USA (CogUSA), another study that is closely related to HRS, has an even more extensive set of cognitive measures. Cognitive measures have also been included in *Understanding Society*, UK Biobank Study, Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS), and Understanding America Study (UAS). The tests included in these studies are listed in Table 1. The mode of administration of the cognitive tests varies across the studies (and sometimes within a study), but all incorporate some form of self-completion.[1]

In general, cognitive assessments in surveys share several qualities that complicate implementation across modes. First, a complete assessment is multidimensional, often including tests of memory, reasoning, orientation, calculation, language, knowledge, and fluid intelligence (see Perlmutter, 1988; Salthouse, 1999). Second, many of the tests employed are relatively time-consuming, and may require special materials or conditions for full administration; for example, show cards, sound capabilities (in web or computer-assisted self-interviewing [CASI] modes), programmed adaptive tests that display different questions based on previous answers, or the

[1] Other resources for self-administered tests include the NIH Toolbox (http://www.healthmeasures.net/explore-measurement-systems/nih-toolbox/intro-to-nih-toolbox/cognition) and Cambridge Cognition (http://www.cambridgecognition.com/products/cognitive-research/web-based-testing). General vocabulary or "Wordsum" cognition tests used to measure verbal crystallized intelligence are also found in the General Social Survey and the American National Election Study (Malholtra, Krosnick, & Haertel, 2007).

ability to measure response time. Finally, detailed instructions or interviewer assistance may be necessary to successfully administer certain tests.

Any attempt to use multiple modes (as most of the studies in Table 1 do) will introduce potential measurement comparability issues. Several recent papers have assessed the role of computerization in cognitive testing for the field as a whole (Wild et al., 2008; Zygouris & Tsolaki, 2013) or for the development of specific tests (Runge, Craig, & Jim, 2015; Ruano et al., 2016). However, a systematic investigation of the factors that influence the measurement of cognition in large-scale surveys is lacking in the field.

Mode Effects and Survey Response

Survey mode is recognized as multidimensional, with effects on numerous aspects of total survey error. In particular, the integration of web surveys raises many considerations that are specific to the issues and approaches of this mode as it evolves (Couper, 2000, 2011). With large-scale longitudinal studies facing pressures to move to the web, however, little empirical work exists detailing how to implement tests on the web. In many cases, these pressures result in mixed-mode formats that may trade off error sources and costs (Dillman, 2000; de Leeuw, 2005; Jäckle, Roberts, & Lynn, 2010).

As Jäckle, Roberts, and Lynn (2010) note, mode effects studies often serve one (and sometimes both) of two purposes. First, a number of studies test for *comparability* of estimates across modes, either considered in reference to a standard or existing mode (as in their discussion of the European Social Survey) or against a "gold standard" or benchmark. These studies may be in the form of testing for data "completeness," differences in marginal effects, effects on overall psychometric properties (such as factor analysis of latent constructs), comparisons of relationships between variables and trajectories by mode, or formal equivalence tests (e.g., Klausch, Hox, & Schouten, 2013; Cernat, 2015a, 2015b; Mariano & Elliott, 2017).

Other studies examine *how* and *why* these effects arise, often using experimental designs to draw conclusions that are narrower in scope. These include specific tests of mode features, such as visual versus auditory processing by controlling the use of show cards (e.g. Jäckle, Roberts, & Lynn, 2010), or of hypotheses for mode differences such as the role of social distance and social desirability in studies comparing interviewer- and self-administration (e.g. Aquilino,

1994; Holbrook, Green, & Krosnick 2003; Kreuter, Presser, & Tourangeau, 2008; Heerwegh & Loosveldt, 2011). While most of these studies involve experiments, others take advantage of studies that reinterview respondents in different modes either by design or given their longitudinal nature (see, e.g., Cernat, Couper, & Ofstedal, 2016; Al Baghal, 2017). We take the latter approach and consider mode effects on measurement, holding the sample source constant and taking advantage of multiple modes of administration within and across waves in a single study.

For the measurement of cognitive ability in a longitudinal study, both comparability and assessing mechanisms are relevant. If comparability across modes and waves of data collection is not achieved, researchers risk drawing inaccurate inferences about both cross-sectional estimates and longitudinal trends. Furthermore, understanding the specific mechanisms behind mode differences can help researchers design cognitive assessments that are comparable across a broader range of contexts. These design decisions will out of necessity be driven not only by theory and knowledge about how modes differ, but by practicality. For example, considerations involved in administering "equivalent" cognitive tests across modes may include the ability to administer tests with sound, which favors interviewer administration for ease of administration; with complex and adaptive programming, which favors computerized administration; and in an accessible manner, which may vary based on the specific design of an instrument.

Cognition in a Multi-Mode Context

A number of features of survey modes have been identified as potentially affecting measurement (see, e.g., Couper, 2011; Jäckle, Roberts, & Lynn, 2010; Tourangeau, Rips, & Rasinski, 2000, chapter 5). Several of these are particularly relevant for the measurement of cognition.

First, modes differ in the presence of an interviewer. Interviewers are associated with higher levels of socially desirable responding, but interviewers also reduce missing data and other errors in surveys. While cognitive ability may not be particularly vulnerable to social desirability influences, respondents might be sufficiently hesitant to perform poorly on a cognitive test that they are unwilling to attempt the test in the presence of an interviewer. Similarly, respondents might be more likely to guess on the web or to admit that they don't know

(if provided the option to do so). The presence of an interviewer may introduce time pressures and increase performance anxiety. On the other hand, from a test comprehension and interpretation perspective, interviewers clearly play an important role in explaining the structure of a cognitive test, providing instructions and ensuring that a test is completed once started. Still, interviewers may also (whether subtly or overtly) assist respondents in the completion of a cognitive performance test, for example by suggesting the correct answer or suggesting respondents try again if an answer is incorrect. Little is known about the behavior of interviewers during cognitive testing and the effect they may have on outcomes.

Furthermore, when completing a cognitive assessment on the web, respondents may feel more comfortable using distributed cognition (see Clark & Chalmers, 1998; Hutchins, 1995) for example, by using aids such as a pencil and paper, a calculator, looking up answers online or consulting others. Estimates of seeking outside help, often considered to be "cheating," range from minimal to pervasive in experimental studies of knowledge questions (e.g. Clifford, & Jerit, 2016; Munzert & Selb, 2017).

Another key dimension on which modes vary is the medium of communication. This includes both the presentation of stimulus material and the delivery of the response. Telephone surveys (for example) restrict the channel of communication to verbal for both. Face-to-face surveys could include visual materials (e.g., show cards). A self-administered (e.g., CASI) component could include both visual and verbal (sound) stimulus material. Online administration can include both visual and verbal stimuli, although (as we discuss later) the mode is predominantly visual in nature. The entry of responses can also differ across modes, from verbal answering in telephone and face-to-face to selection of options or typing in CASI and web. But on a finer level, even the direct selection of an object using a touchscreen (e.g., on a tablet) may change performance on a task relative to using a mouse and pointer, depending on the dexterity of respondents and their familiarity with the technology. This is especially true if speed is an element of the test. In other words, the mode of administration may change the nature of the test.

Time pressures present competing potential influences on data quality. Time pressures that are likely present in the interviewer-administered modes, and especially on the telephone, may lead to worse performance or missing data compared to a self-administered survey if respondents take the opportunity to think further about the questions on the web. In contrast,

respondent focus might be lessened without the presence of the interviewer; and if satisficing (Krosnick, 1991) is more prevalent in the self-administered cognitive tests than in the interviewer-administered cognitive tests (Heerwegh, 2009), more missing data, speeding, or straightlining could result on the web.

Visual presentation and processing may not only contribute to primacy effects as opposed to recency effects (Schwarz, Hippler, & Noelle-Neumann, 1992), but could also fundamentally change the cognitive task at hand—for example, from a recall task to a recognition task, with the latter generally considered to be an easier cognitive task. The effects of cognitive burden may be fairly complex. For example, in a randomized mode experiment, Chang and Krosnick (2010) found that respondents with lower cognition levels responded with higher levels of concurrent validity to a set of questions on the web than when the survey was administered over an intercom by an interviewer, suggesting that the demands posed by working memory might yield lower response quality among those at the lower end of a range of cognitive abilities in auditory modes.

Mode effects may vary across subgroups of the population, and may be especially problematic given the importance of measuring cognitive decline at the low end of the spectrum. Respondents who are less computer literate, or who suffer from physical or cognitive limitations that may make it difficult to use a computer, may be particularly susceptible to mode effects when the web is utilized. Cognitive tests must thus be designed in a manner accessible to those with disabilities, health problems, visual or aural impairments, or low levels of literacy, cognition, and computer skills, if they are to be considered equivalent across modes.

Existing Mode Comparisons of Cognitive Ability

Despite the fact that cognitive measures are widely used in single- and multi-mode surveys and that they have the potential to be subject to a variety of mode effects, few mode comparisons of cognitive measures have been carried out that compare interviewer and web administration. Those that have been conducted are limited to individual measures and specific contexts.

While comparisons between different interviewer-administered modes exist, they are also limited. Rodgers, Ofstedal, and Herzog (2003) found higher scores on the telephone than via face-to-face interviewing, but acknowledge that mode may be confounded with cognitive ability

due to lack of clean random assignment. An earlier randomized mode experiment in HRS/AHEAD found no differences in average cognitive scores between telephone and face-to-face interviewing (Herzog et al., 1999).

More recent comparisons including web administration are rare. In one such study, Runge, Craig, and Jim (2015) compared the performance of female respondents from several waves of the HRS to an independent sample of women who participated in the 2013 Women's Health Valuation (WHV) study. WHV participants were recruited from an existing online panel of U.S. adults and sampled using quotas for age and race/ethnicity. The WHV replicated the HRS immediate and delayed recall tasks on the web. The authors found that WHV respondents had higher immediate and delayed recall scores than HRS respondents from equivalent years. Key predictors of recall were largely similar across studies, although there were a few differences. For example, in the WHV, those with poor self-reported memory recalled fewer words than similar HRS counterparts. The authors note the potential for literacy level and typing skills, the opportunity for interviewer administration to slow the task for respondents low in cognition, and the different cognitive processes for visual versus auditory administration of the task to explain the differences.

A more controlled mode comparison of one cognitive measure was carried out by Gooch (2015) in a mode experiment. Visitors to a television research facility were randomized after agreement to participate in either a face-to-face interview or a web survey that took place on the facility's computers. Among other measures, respondents were asked to respond to the Gallup–Thorndike Verbal Intelligence Test, also called the "Wordsum" test. Gooch found that the modest-to-difficult questions in the battery were answered correctly more often on the web than in the face-to-face survey, though the pattern reversed for easier questions (they were answered correctly more often in face-to-face administration). However, the ordering of points in an item response theory model was identical across modes and the mode effects did not yield significant differences when used in multivariate substantive models.

More recently, Al Baghal (2017) conducted an analysis of cognitive measures in the *Understanding Society* Innovation Panel (IP7, conducted in 2014), a mixed-mode design involving face-to-face interviewer administration and web administration. In the face-to-face mode, the cognitive measures were self-administered as part of the CASI module. The measures

he examined included the number series, verbal analogies, numeracy, and four forward digit-span tasks, testing working memory capacity. Cases were randomly assigned to a sequential mixed-mode protocol (web then face-to-face) or face-to-face only. Among other differences, those responding via the web in the mixed-mode design were significantly younger, more educated, and more likely to be daily (as opposed to less frequent) Internet users. To control for self-selection, Al Baghal examined responses to the cognitive measures in IP1 (where all were interviewed in-person) and used inverse probability of treatment weighting (IPTW) methods to account for differential selection into modes. He found that web respondents perform significantly better than face-to-face respondents in both the measures of inductive reasoning, numeracy and recall (measured by forward digit span). The study maintained a consistent visual presentation across all modes, so differences are not attributable to aural versus visual modes of delivery.

Given the widespread use of cognitive measures in both methodological and substantive models, there is need for more systematic mode comparisons to evaluate the ways in which survey mode may affect the usefulness of the measures and their validity with respect to an individual's true cognitive state. We do this in the current study by examining measurement differences for five cognitive tests. We restrict our analysis to a set of respondents who completed the same measures in both interviewer- and self-administered modes in order to distinguish measurement effects from any potential selection effects, especially those related to cognitive ability.

DATA AND METHODS

Data Source

The Health & Retirement Study (HRS) is a longitudinal study of older adults in the United States conducted by the University of Michigan. It surveys respondents over the age of 50 and their spouses, with successive cohorts of respondents added over time to maintain a representative sample of the study population. Respondents participate in core surveys every other year, with individuals under the age of 80 randomized to receive a face-to-face or telephone interview in alternate waves. The content of the telephone and in-person interviews are essentially identical, with the exception of added physical measures, biomarker collection, a

psychosocial leave-behind questionnaire and data linkage consent requests in the face-to-face administration. In years that do not contain a core interview, the study team fields a variety of off-year efforts, including web and mail surveys. The web surveys, which were fielded in alternate years between 2003 and 2013, contain some questions from the core interview, as well as a range of new topics. (See the HRS website for more information: http://hrs.isr.umich.edu/.)

Analytic Sample

Our analysis utilizes data from 4,223 respondents between the ages of 50 and 80 who self-completed (i.e., did not have a proxy respondent) each of the 2012 core, 2013 web, and 2014 core survey requests; and who did so in the mode they were assigned according to the mode randomization utilized between 2012 and 2014. The 2013 web survey was administered to a sub-sample of HRS participants who reported in their most recent core interview that they had access to the Internet. A random 80% of those with Internet access were selected for the 2013 web sample (n=7,744) and 75% of those sampled completed the survey (n=5,813). We remove those younger than 50 (age-ineligible spouses of HRS sample members) as well as those who completed their core interview by proxy (for whom the standard cognitive tests were not administered). In order to cleanly perform between-sample comparisons, and to leverage the benefits of randomization, we remove respondents older than 80 (who are nearly always assigned to face-to-face data collection) as well as those who did not complete the 2012 and/or 2014 core interview in the mode to which they were assigned.[2] Table 2 displays the breakdown of the sample starting with the 2012 core interview, with successive restrictions imposed so that each member of the final sample (n=4,223) had completed the 2012, 2013, and 2014 interviews in the assigned mode.

Our focus is on investigating within-respondent mode effects, rather than generalizing to a broader population. However, given the restrictions we impose, it is important to recognize the demographic differences from the general HRS cohort that exist. The analytic sample is comprised of individuals who have consistently responded to survey requests, who have Internet access and are willing to respond to a web survey. Socio-demographic comparisons of the analytic sample and the broader HRS 2012 core sample show that the analytic sample is

[2] Most respondents in the analytic sample completed the core survey in the assigned mode: 96% in 2012, 95% in 2014.

disproportionately white, educated, employed, healthy, young, and has higher average income (Table 3). The analytic sample is also higher functioning with respect to cognition than the full HRS sample. This latter result suggests that any mode differences in cognition that we observe in the analytic sample may be conservative relative to a more representative sample that has a broader range of cognitive ability.

Administration of Cognitive Tests

Since the primary core modes of HRS are telephone and face-to-face, the main considerations for test selection have been to include items that can be administered over the phone (that is, not reliant on visual aids) and to keep the cognitive battery sufficiently short to minimize respondent burden. Originally, tests were primarily drawn from the Telephone Interview for Cognitive Status (TICS) screen, based on the Mini-Mental State Exam (Brandt, Spencer, & Folstein, 1988). HRS cognitive measures were expanded starting in 2010 to provide more differentiation at the higher-end of functioning (Fisher, McArdle, McCammon, Sonnega, & Weir, 2013). Five tests that were administered in 2012 core, 2013 web and 2014 core are utilized in the present analysis.

The ***quantitative number series*** measures quantitative reasoning and fluid intelligence, and is a six-item, block adaptive test based on answers to 6 out of 15 possible items. All respondents start with the same three items, but the difficulty of the items shown in the second set of three items depends on the respondent's answers to the first set. Respondents are asked to fill in the blank in a series—"for example, if I said the numbers '1 2 BLANK 4,' then what number would go in the blank?" We utilize the "W-scores" in the HRS dataset. These are standardized scores designed to be comparable to the Woodcock-Johnson III (WJ-III) test battery on which this task was based (Fisher, McArdle, McCammon, Sonnega, & Weir, 2013). The score ranges from 409 to 569. A 10-point decrease in the score represents halving the probability that a respondent answers a given item (or one of equal difficulty) correctly. Notably, and as a departure from other tests, the respondent is asked to write down the series of numbers in the interviewer administered modes before communicating the answer to the interviewer[3], introducing visual processing into the telephone and face-to-face contexts. The number series

[3] We have no indicator of the extent to which respondents complied with this instruction.

test is administered in alternate waves in the core interview (2012, 2016, etc.). It was administered to a random subsample of participants in the 2013 web survey.

Numeracy is measured via respondents' answers to three questions developed by Lipkus, Samsa, and Rimer (2001; see also Huppert, Gardener, & McWilliams, 2004): how many individuals will be expected to get a disease out of 1,000 given a 10% chance ("chance of disease"); how much money will be received by each of five winners for a $2 million dollar lottery prize ("lottery split"); and how much money will result after two years if a sum of $200 yields 10% interest per year ("compound interest"). In the standard interviewer administration, this last item is only asked if either of the first two is answered correctly. In the web survey, a more restrictive rule was applied that asked the third question only if *both* of the first two questions were answered correctly. To eliminate this confound of mode and test administration, we restrict our analysis to those respondents who would have received the same treatment in both modes (i.e., we exclude respondents who answered only one of the first two questions correctly). (We obtain similar results when applying an alternative assumption, that respondents who answered only one of the first questions correctly in the FTF or telephone administrations would have answered the third wrong.) We follow the 0-4 scoring of Levy et al. (2014) in which partial credit is given for a "nearly correct" answer to the compound interest question (i.e. $240 garners one point; a correct answer of $242 garners two points). The numeracy items are administered in alternate waves in the core interview (2010, 2014, etc.). They were administered to the full 2013 web sample.

The ***serial 7s*** test measures working memory (Ofstedal, Fisher, & Herzog, 2005) and requires respondents to subtract 7 from 100 five consecutive times, with the composite score ranging from 0-5 (representing the number of correct answers). Respondents are given full credit for later correct answers regardless of what their first answer was (e.g., a second response of 83 would be counted as correct if the first response was 90, even though 90 is incorrect as the first response). The serial 7s test is administered in every wave of the core interview and was administered to the full 2013 web sample.

The ***verbal analogies*** test is administered in similar fashion to the quantitative number series and measures verbal reasoning. The six-item test is block-adaptive and administered from a set of 15 possible items, with the difficulty of the second set of items dependent on the

respondent's answers to the first set. For this test, respondents are asked to fill in the missing word in an analogy such as "Mother is to Daughter as Father is to…" The standardized W-score ranges from 435 to 560. The verbal analogies test was administered to a small random subsample in the 2012 core interview and to a random half sample in the 2013 web survey. Starting in 2014, verbal analogies is administered to the full core sample in alternate waves (2014, 2018, etc.).

Word recall is administered via both immediate and delayed tests in the interviewer-administered waves of the HRS. For this pair of tests, the interviewer reads one of four randomly selected lists of ten words. Respondents are then asked to immediately recall the words remembered, as well as recall the same words after a delay of 2-3 minutes (after administration of other questions). For the web survey, a word recognition test was used in place of word recall. In the word recognition test, respondents first listened to an audio recording of ten words being read (using the same lists as for interviewer administered) and were then immediately presented with a list of 20 words and asked to select the ones from the recorded list. No delayed recall/recognition test was administered on the web. The recognition test required respondents to enable sound on their computer and pass a "sound test" in order to hear the series of words read aloud. Only 39% (n=982) of respondents selected for the test were able to successfully play sound on their computers.

Due to the alternate wave and subsampling design for the cognitive measures in the core waves and 2013 web survey, the resulting analytic sample size permitting within-respondent analysis across waves varies substantially across tests. Table 4 shows the analytic sample sizes for each of five cognitive tests.

Methods

All of the analyses presented are unweighted and do not take the complex survey design into consideration, as we are purposefully working with a small subset of cases that are not representative of the entire HRS sample. We describe each set of analyses in turn below.

Item Missing Data

Given the varied nature of the cognitive tests used both in number of items and question design, an overall assessment of item-missing data across modes requires definitions of

"missing" values that are specific to each test. Additionally, since no explicit "don't know" response was provided to web survey respondents, we are unable to determine whether missing data on the web survey was due to the respondent not knowing the answer, refusing to answer the question, or for some other reason. For the number series and verbal analogies, we define missing based on the first item for each test. If the respondent skipped the first item (in web) or responded either "Don't know" or "Refused" (in interviewer mode), that respondent is coded as missing for that test. If the respondent answered the first item but had missing data on a subsequent item, they are not coded as missing and are assigned a score based on the completed items. For serial 7s, standard coding for the test assigns a "Don't Know" response to a score of 0 correct. We follow this scoring convention for both modes; given the inability to distinguish types of missing responses on the web, however, for the purposes of item-missing data comparison we count "Don't know" answers as item-missing. Similarly, on interviewer-administered word recall measures responses of "None remembered" and "Refused" are treated as item-missing for comparability with a skipped item on the web, though also assigned a score of 0. For numeracy, we consider answers to the individual questions separately according to the same guidelines. We conduct *paired t-tests* when examining differences in levels of item missing data between interviewer-administered waves (telephone and face-to-face) and web administration (which are based on the same individuals) and *between-sample t-tests* when examining differences between the telephone and face-to-face modes within a wave.

Completion Time

We use between- and within-sample t-tests to compare completion times for each of the cognitive tests. Times were summed from the raw paradata at the page-visit or field-visit level to the test level.[4] Outlier times beyond the 95th percentile at the test level in a given wave are top-coded to the 95th percentile and comparisons are restricted to cases with a positive time spent on the test in all three waves. For all tests, time to read (or be read) introductions and practice questions was counted as part of the test administration time. For word recognition in the web

[4] For select cases where paradata were not available, already-summed test-level times from the public use dataset were substituted if a suitable variable could be identified. In most cases, these matched the manually summed times; however, for several tests the manually summed times were consistently 2-3 seconds different from the times presented in the public use dataset. Such differences affect all cases in 2014 equally. For an additional small number of cases, timings could not be linked; thus sample sizes vary slightly from overall analyses.

survey, time to carry out the "sound test" was included in the test-level time, given the integral nature of this task to carrying out the word recognition test.

Overall Differences in Scores

We again use between and within-sample t tests when comparing mean scores for each of the cognitive tests across modes. To keep the sample size and composition constant across comparisons, only cases with a substantive answer for the test (including 0, where relevant) in all three waves are included; that is, any case with an overall missing score (due to a refusal, don't know response, or leaving the test item(s) blank) on any wave is excluded.

Correlations between Measures

Because the subsampling design described previously precludes most confirmatory factor analysis that could assess the consistency of the factor structure of the cognitive tests across modes, we focus on correlations between *pairs* of tests within a wave, as well as correlations for a given test across waves. Again, we restrict the sample to those who substantively answered the relevant tests in all waves examined in order to keep sample size and composition constant.

Trajectories over Time

As Rabbitt, Diggle, Holland, & McInness (2004) summarize, many longitudinal studies of cognition have one of three aims—to assess trajectories of change, and particularly whether cognitive decline accelerates in old age (e.g. Hertzog & Schaie, 1988); to assess how rates of change differ between baseline mental abilities; and to assess whether these trajectories are affected by a wide range of demographic, social, and environmental factors. We provide an initial assessment of change between paired waves to assess whether a given respondent performed better, worse, or about the same between the waves, where change is defined as approximately a one-standard deviation difference from the previous wave. For word recall and recognition, this is defined as a change of two or more words recognized; for the verbal analogies score, it is defined as a change of 24 points or more on the standardized scale; for the number series, a change of 26 points or more on the standardized scale; for serial 7s, a change of one or more of five possible responses correct; and for numeracy, a change of one or more points on the four-point scale.

Two types of longitudinal models are subsequently used to test the hypothesis that the trend over time varies with mode. First, a random intercept repeated measures model with an autoregressive error structure is fit to observations clustered by respondent, controlling for whether or not the individual began in 2012 with face-to-face or telephone administration. This multilevel modeling approach is commonly applied to repeated measures data and, when it incorporates random coefficients or "slopes" as well with larger numbers of time points, parallels the use of latent growth curve models to assess individual-level trajectories (for one discussion of their similarities, see Hox & Stoel, 2005). We fit a simpler model here given only three time points, though further analysis could estimate individual trajectories. Second, a latent class growth analysis (Nagin, Jones, Lima Passos, & Tremblay, 2016), similar to the growth mixture modeling approach, is fit to the same data to assess whether respondents appear to fall into different classes of trajectories over time. This model tests the hypothesis that latent groups of respondents can be defined by assessing differences in their trajectories in cognitive scores (e.g., examining the shape of trajectories and one's probability of following a given trajectory), specifying various trajectory forms and class solutions in model selection.

Models Predicting Cognition as an Outcome

Finally, we compare results from models that utilize a set of covariates to predict cognitive ability across tests and mode of administration. Predictors of cognition were drawn from relevant substantive literature (e.g., Rodgers, Ofstedal, & Herzog, 2013; Langa et al., 2017) and include age, gender, race/ethnicity, education, work status, chronic conditions, depressive symptoms, household income, and a count of types of internet activities the respondent reported (e.g., financial, other commerce, social network, contact (e.g. email), news/entertainment, and/or work-related tasks). We use OLS regression models for this analysis.

RESULTS

Item-Missing Data

First, we assess the completeness of the data across modes. Table 5 presents the percentage of respondents with item-missing data on each test, using the criteria described earlier. The trend in item-missing data varies by the type of test. Verbal analogies and number series, the two longer, adaptive tests, have higher rates of missing data on the web than in the interviewer-administered modes. However, numeracy and serial 7s yield *less* item-missing data on the web. While word recognition on the web yields a substantially higher percentage of item-missing data than in interviewer-administered modes, the web version did not include a "none remembered" option; thus, we cannot disentangle item-missing data from these responses as in the interviewer-administered modes. There are no significant differences in item missing data between telephone and face-to-face modes.

Completion Time

An analysis of completion time for the cognitive tests across modes suggests that all of the tests except the number series take *longer* to administer on the web—going against conventional wisdom that web administration might be more efficient than interviewer administration (Table 6). As the sole exception, the number series is administered more quickly on the web than via the telephone, though it takes about the same amount of time as in face-to-face administration. No significant differences are observed between telephone and face-to-face administration.

Differences in Mean Scores

Table 7 illustrates the differences in mean cognitive scores across modes and waves for the analytic sample, utilizing for each comparison only cases for which a substantive response was given. In all cases, within-subject tests suggest that mean scores are significantly higher for web than for telephone and face-to-face (all within-subject differences are significant at $p < .001$). While most between-subject comparisons for telephone versus face-to-face in a given wave yield no significant differences by mode, both verbal analogies and word recall exhibit significantly higher scores via face-to-face administration than telephone administration ($p <$

.05), but only in 2014. Overall, the pattern suggests that respondents do better on cognitive tests on the web, though in some cases differences are small and may not be substantively important.

We also examined the percentage of respondents achieving the *maximum score* for each test and found that this percentage is substantially higher on the web than in either telephone or face-to-face interviews (Table 8). These differences are particularly large for word recall/recognition: 27% of respondents correctly recognized all ten words on the web, as compared to no more than 2% recalling all ten words correctly in any other mode or wave. This difference may be the result of several factors: lower working memory demands in a visual mode, similar to the findings of Chang & Krosnick (2010); recognition being a cognitively easier task than recall; or respondents writing words down or otherwise using aids to remember the words via self-administration. Regardless of the underlying reason, it is clear that web administration yields higher performance than face-to-face or telephone administration for these tests.

Correlations between Measures

Results shown in Table 9 suggest that cognitive test scores are more weakly correlated with each other when web administration is involved, suggesting a potential decline in reliability when modes are mixed. Where tests are available across all three waves of data collection, correlations between administrations of a single test in interviewer-administered waves are higher than the correlations between web and interviewer administration, despite the fact that the latter administrations are closer in time than the former. We also examined between-test correlations and correlations for each test with self-rated memory *within* a wave (Table 10). For all tests except the number series, correlations are uniformly lower in the 2013 web administration than in the other waves and modes.

Trajectories over Time

Figure 1 displays the distribution of respondents performing worse (left portion of bar), about the same (middle), and better (right) from one wave to the next, where change is defined as a one standard deviation difference. Changes in either direction are relatively evenly distributed when examining the two interviewer-administered waves at the top of the figure. *More* change is observed overall for the one-year intervals involving switches between interviewer- and self-

administration, and the direction of this change appears to diverge by mode sequence. Respondents tend to do better in a subsequent wave when moving from interviewer- to self-administration (middle panel); and worse in a subsequent wave when moving from self- to interviewer-administration (bottom panel).

To further address these trajectories over time, we estimate several longitudinal models. First, we fit a random effects model with random intercepts for each respondent, a nonlinear fixed effect of time, and an autoregressive error structure. Results from this model suggest a potential curvilinear pattern over time for each cognitive outcome (Table 11). While there is a significant, positive fixed effect for the 2013 time point as compared to 2012, there is no such effect for the 2014 time point as compared to 2012 for word recall and serial 7s. In other words, while scores for the interviewer-administered waves do not differ with time, changing from interviewer to self-administration is associated with an increase in cognitive score. While the verbal analogies score does display significant increases in both years, the increases are substantively quite small (especially between 2012 and 2014) and may not be practically significant. We estimated the same random effects model with a set of sociodemographic and health covariates (age, gender, race, education, chronic conditions, count of internet activities; results not shown) and observed the same pattern of fixed effects of time across waves.

Second, we ran a set of latent class growth models, specifying quadratic trajectories for each class, and testing solutions with one to four classes. The Bayesian Information Criterion indicated a three class solution as the best fit for each cognitive outcome, with results shown in Figures 2-4. For word recognition/recall, 89% of respondents fall into one of two classes that matches the curvilinear trajectory described above, with higher scores on the web (Figure 2). For serial 7s, only 15% of respondents fall into a class that suggests a practically significant curvilinear trajectory; the other two classes showed essentially no change across the three waves (Figure 3). Finally, the three-class solution for the verbal analogies score shows a small increase in scores between 2012 and 2013, followed by a decrease in 2014 for one class, predicted to represent 44% of respondent trajectories (Figure 4).

For both the random effects and latent class growth models, whether a respondent completes the survey via telephone or face-to-face in 2012 does not predict which class of trajectories he or she has the highest probability of falling into.

One alternative explanation for these patterns of results is that of a practice effect. Other literature examining longer time trajectories has found an increase in score from T0 to T1 followed by slight decreases thereafter, suggesting that cognitive decline is masked somewhat by the presence of the practice effect (Jacqmin-Gadda, Fabrigoule, Commenges, & Dartigues, 1997; Unger, Belle, & Heyman, 1999; McArdle, Fisher, & Kadlec, 2007). However, because most of these cognitive tests have been administered in HRS prior to 2012, this explanation has less relevance here. The only test that could potentially be subject to a practice effect is verbal analogies, which was first administered in 2012.

Substantive Models

Finally, we estimated a set of OLS regression models to assess whether predictors of cognitive scores were consistent across modes. Results from these models yield somewhat inconsistent results, with many of the predictors displaying different associations across modes (Table 12). For example, education is positively related to word recall in the interviewer-administered waves, but it is unrelated to word recognition on the web. This same pattern by education is found for Serial 7s. Women had higher verbal scores than men in the web administration, but no differences were observed for the interviewer-administered waves. In contrast, women scored better than men on the interviewer-administered word recall, but there was no gender difference in word recognition scores on the web. Hispanics performed significantly worse than non-Hispanic Whites on Serial 7s in the interviewer-administered waves, but there was no difference in the web administration. Of note, scores on all tests except word recall/recognition are positively related to Internet usage.

DISCUSSION

As surveys are increasingly being pushed to the web in order to manage costs, this shift raises concerns about selection effects and measurement comparability across modes. This is especially concerning for longitudinal surveys, for which a shift in mode during the course of the survey could disrupt estimates of time series for key measures. As documented in Table 1, the measurement of cognitive functioning is a key domain in longitudinal studies of health and aging. Cognitive measures may be particularly susceptible to mode effects, because many of the measures that were developed for interviewer administration (particularly by telephone) may not

be easily adapted for self-administration. We address this topic using a variety of cognitive measures administered in the Health and Retirement Study.

Most studies of mode effects are based on comparisons of different respondents. Although often randomized to mode, there is selection into who responds in the assigned mode that can confound comparisons across modes. A major strength of our study is that the same individuals received the same (or similar) cognitive tests in different modes, which allows us to control for selection effects. A limitation is that the analytic sample is different from the full sample in important ways and results may not be generalizable to the general population, which is characterized by a broader range of cognitive ability. Another limitation is the small sample sizes for some of the comparisons.

Our results suggest that survey mode *does* affect estimates of cognitive ability. The main differences are between web vs. interviewer-administered modes (telephone and face-to-face), although we also observe some differences between telephone and face-to-face administration. For all of the cognitive measures, respondents performed better on the web than in either interviewer administered mode. This was true regardless of whether the web administration occurred before or after the interviewer administration. As a result, measures of trajectories over the three waves covered in our study are adversely affected. A sizeable proportion of respondents are characterized by an inverse U-shaped trajectory, whereby cognitive performance increases between T1 (interviewer) and T2 (web) and declines between T2 and T3 (interviewer). This pattern, which is contrary to the expected pattern of age-related cognitive decline, is especially apparent for word recall/recognition and verbal analogies.

Whereas performance on the cognitive measures is consistently higher on the web than interviewer modes, correlations between pairs of tests or within-test correlations across waves are consistently lower for measures administered on the web. Likewise, associations between some key predictors of cognition (e.g., age, sex, education) differ across modes. Thus, mode appears to affect not only levels of cognitive performance, but also other properties of the cognitive measures.

It is unclear which mode yields more valid results. On the one hand, respondents may feel more anxiety or pressure with an interviewer present and perform worse than their true

ability. On the other hand, without an interviewer to observe, web respondents may use aids (e.g., calculator, online searches, write down words, etc.) and/or take more time to think about their answers when completing the tests and perform better than their true ability. Al Baghal's (2017) findings of differences between CASI and web administration for identical tests and our finding that HRS respondents take longer to complete most of the tests on the web than in interviewer modes suggest that either or both of these circumstances may be at play. Satisficing (or sub-optimal responding) is another potential confounder, and on this our evidence is mixed. Our findings of higher performance and longer completion times on the web seem to suggest less satisficing; however, higher missing data rates on the web suggest more satisficing.

Regardless of which mode yields more valid results, however, our findings suggest that a switch from interviewer to web-based administration (or vice versa) will affect measures of cognitive performance. If the mode differences were limited to levels and occurred uniformly across respondents, then a simple calibration or control for mode may suffice. However, because we find mode differences in associations—both between pairs of cognitive tests and between cognitive performance and known predictors of cognition—this suggests that a more nuanced approach will be needed. Given the nature of our analytic sample—who are disproportionately white, educated, employed, healthy, young, with higher average income, and higher in cognitive functioning than the full HRS sample—it is possible that our results are conservative and that those with lower cognitive abilities would, if pushed to the web, display even larger mode effects.

These results provide lessons for HRS and other ongoing studies that plan to add web as a new mode, as well as for new multi-mode studies that are getting underway. The first is to be clear about priorities. The need to maximize comparability across modes calls for a very different design approach than the desire to take maximum advantage of the capabilities of individual modes. For example, self-administration allows for the use of visual stimuli, which is not possible in telephone interviews. On the other hand, tests that require verbal communication (reading or repeating words, counting backwards, naming animals) may be difficult or impossible to replicate in a web survey where it is not practical (or ethical) to control settings on respondents' computers. If it is not critical for the study's purposes to administer the same measures in all modes, then using multiple modes could enhance the measurement of cognition

by allowing for a broader set of measures than would be possible in a single mode. However, for most studies, this is likely not an option, or at least not the top priority.

Assuming that comparability of tests is a priority, it is important to identify tests that are suitable for administration across different modes. HRS's attempt to use sound capability to administer the word list in the 2013 web survey was instructive in this regard, in that less than half of respondents reported having sound capability enabled on their computer. Efforts to test video stimulus material on the web have obtained similar results (Mendelson, Gibson, & Romano-Bergstrom, 2017). This means that, at least for the older population in the U.S., tests that have a verbal/audio requirement are not advised for web administration. As technology continues to develop it may become easier to use comparable protocols across modes.

Regardless of how comparable the tests are in terms of administration protocol, however, there are still likely to be some measurement differences by mode. At a minimum, this means that careful attention to and calibration of the data will be needed. To the extent possible, experiments that are designed to test for mode effects and inform calibration would be valuable to incorporate in longitudinal studies administering cognitive measures.

ACKNOWLEDGEMENTS

Data used in the study come from the Health and Retirement Study, which is funded by a grant from the National Institute on Aging (U01 AG009740) with supplemental support from the Social Security Administration. HRS is conducted by the University of Michigan.

REFERENCES

Al Baghal, T. (2017). The effect of online and mixed-mode measurement of cognitive ability. *Social Science Computer Review*, 1-15. (Published online December 21, 2017.)

Aquilino, W. S. (1994). Interview mode effects in surveys of drug and alcohol use: A field experiment. *Public Opinion Quarterly, 58*(2), 210-240.

Brandt, J., Spencer, M., & Folstein, M. (1988). The telephone interview for cognitive status. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology, 1*(2), 111-117.

Cernat, A. (2015a). The impact of mixing modes on reliability in longitudinal studies. *Sociological Methods & Research, 44*(3), 427-457.

Cernat, A. (2015b). Impact of mixed modes on measurement errors and estimates of change in panel data. *Survey Research Methods*, 9(2) 83-99.

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.

Chang, L., & Krosnick, J. A. (2009). National surveys via RDD telephone interviewing versus the Internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly*, *73*(4), 641-678.

Chang, L., & Krosnick, J. A. (2010). Comparing oral interviewing with self-administered computerized Questionnaires: An experiment. *Public Opinion Quarterly*, *74*(1), 154-167.

Clark, A., and Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), pp. 7-19.

Clifford, S., & Jerit, J. (2016). Cheating on political knowledge questions in online surveys: An assessment of the problem and solutions. *Public Opinion Quarterly*, *80*(4), 858-887.

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *The Public Opinion Quarterly*, *64*(4), 464-494.

Couper, M. P. (2011). The future of modes of data collection. *Public Opinion Quarterly, 75*(5), 889–908.

De Leeuw, D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics, 21*(2), 233.

Dillman, D. A. (2000). Mail and internet surveys: The total design method. *New York: Wiley*.

Dillman, D. A., & Christian, L. M. (2005). Survey mode as a source of instability in responses across surveys. *Field Methods, 17*(1), 30-52.

Duffy, B., Smith, K., Terhanian, G., & Bremer, J. (2005). Comparing data from online and face-to-face surveys. *International Journal of Market Research, 47*(6), 615.

Fisher, G.G., McArdle, J.J., McCammon, R.J., Sonnega, A., & Weir, D.R. (2013). New measures of fluid intelligence in the HRS. Ann Arbor, Michigan: Institute for Social Research, University of Michigan, https://hrs.isr.umich.edu/sites/default/files/biblio/dr-027b.pdf.

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly*, *69*(3), 370-392.

Gooch, A. (2015). Measurements of cognitive skill by survey mode: Marginal differences and scaling similarities. *Research & Politics, 2*(3), 1-11.

Heerwegh, D. (2009). Mode differences between face-to-face and web surveys: an experimental investigation of data quality and social desirability effects. *International Journal of Public Opinion Research, 21*(1), 111-121.

Heerwegh, D., & Loosveldt, G. (2011). Assessing mode effects in a national crime victimization survey using structural equation models: Social desirability bias and acquiescence. *Journal of Official Statistics, 27*(1), 49–63.

Hertzog, C., & Schaie, K. W. (1986). Stability and change in adult intelligence: I. Analysis of longitudinal covariance structures. *Psychology and Aging*, *1*(2), 159.

Herzog, A. R., Rodgers, W. L., Schwarz, N., Park, D. C., Knauper, B., & Sudman, S. (1999). Cognitive performance measures in survey research on older adults. In Schwarz, N., D. Park, B.

Knauper, & S. Sudman (Eds.), *Cognition, aging, and self-reports* (pp. 327-340). Psychology Press: Philadelphia, PA.

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, *67*(1), 79-125.

Hox, J., & Stoel, R. D. (2005). Multilevel and SEM approaches to growth curve modeling. In Everitt, B.S., & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (pp. 1296-1305). Chichester: John Wiley & Sons.

Huppert, F. A., Gardener, E., & McWilliams, B. (2004). Cognitive function. In J. Banks, E. Breeze, C. Lessof, & J. Nazroo (Eds.), *Retirement, health and relationships of the older population in England: the 2004 English Longitudinal Study of Ageing* (pp. 217-242). Institute for Fiscal Studies, London.

Hutchins, E. (1995. *Cognition in the wild*. Cambridge, MA: MIT Press.

Jäckle, A., Roberts, C., & Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review*, *78*(1), 3-20.

Jacqmin-Gadda, H., Fabrigoule, C., Commenges, D., & Dartigues, J. F. (1997). A 5-year longitudinal study of the Mini-Mental State Examination in normal aging. *American Journal of Epidemiology, 145*(6), 498-506.

Klausch, T., Hox, J. J., & Schouten, B. (2013). Measurement effects of survey mode on the equivalence of attitudinal rating scale questions. *Sociological Methods & Research*, *42*(3), 227-263.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and Web surveys the effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*(5), 847-865.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213-236.

Langa, K. M., Larson, E. B., Crimmins, E. M., Faul, J. D., Levine, D. A., Kabeto, M. U., & Weir, D. R. (2017). A comparison of the prevalence of dementia in the United States in 2000 and 2012. *JAMA Internal Medicine, 177*(1), 51-58.

Levy, H., Ubel, P. A., Dillard, A. J., Weir, D. R., & Fagerlin, A. (2014). Health numeracy: the importance of domain in assessing numeracy. *Medical Decision Making, 34*(1), 107-115.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, *21*(1), 37-44.

Mariano, L. T., & Elliott, M. N. (2017). An Item Response Theory Approach to Estimating Survey Mode Effects: Analysis of Data from a Randomized Mode Experiment. *Journal of Survey Statistics and Methodology*, *5*(2), 233-253.

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.

Mendelson, J., Gibson, J.L., and Romano-Bergstrom, J. (2017). Displaying videos in web surveys: Implications for complete viewing and survey responses. *Social Science Computer Review*, 35 (5): 654-665.

Munzert, S., & Selb, P. (2017). Measuring political knowledge in web-based surveys: An experimental validation of visual versus verbal instruments. *Social Science Computer Review*, *35*(2), 167-183

Nagin, D. S., Jones, B. L., Lima Passos, V., & Tremblay, R. E. (2016). Group-based multi-trajectory modeling. *Statistical Methods in Medical Research*. Advance online publication. doi: 10.1177/0962280216673085.

Ofstedal, M. B., Fisher, G. G., & Herzog, A. R. (2005). *Documentation of cognitive functioning measures in the Health and Retirement Study.* Ann Arbor, MI: University of Michigan, 10.

Perlmutter, M. (1988). Cognitive potential throughout life. In J. E. Birren & V. L. Bengtson (Eds.), *Emergent theories of Aging* (pp. 247-268). New York: Springer.

Rabbitt, P., Diggle, P., Holland, F., & McInnes, L. (2004). Practice and drop-out effects during a 17-year longitudinal study of cognitive aging. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *59*(2), P84-P97.

Rodgers, W. L., Ofstedal, M. B., & Herzog, A. R. (2003). Trends in scores on tests of cognitive ability in the elderly US population, 1993–2000. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *58*(6), S338-S346.

Ruano, L., Sousa, A., Severo, M., Alves, I., Colunas, M., Barreto, R., ... & Lunet, N. (2016). Development of a self-administered web-based test for longitudinal cognitive assessment. *Scientific Reports*, 6. Retrieved from https://www.nature.com/articles/srep19114.

Runge, S. K., Craig, B. M., & Jim, H. S. (2015). Word recall: Cognitive performance within Internet surveys. *JMIR Mental Health*, *2*(2). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4607399/.

Salthouse, T. A. (1999). Theories of cognition. In Silverstein, M., V. L. Bengtson, N. Putney, &, D. Gans (Eds.), *Handbook of theories of aging* (pp. 196-208).

Schwarz, N., Hippler, H. J., & Noelle-Neumann, E. (1992). A cognitive model of response-order effects in survey measurement. In Schwarz, N., & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 187-201). Springer New York.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press.

Unger, J. M., Belle, G., & Heyman, A. (1999). Cross-Sectional Versus Longitudinal Estimates of Cognitive Change in Nondemented Older People: A CERAD Study. *Journal of the American Geriatrics Society*, *47*(5), 559-563.

Wild, K., Howieson, D., Webbe, F., Seelye, A., & Kaye, J. (2008). Status of computerized cognitive testing in aging: a systematic review. *Alzheimer's & Dementia*, *4*(6), 428-437.

Zygouris, S., & Tsolaki, M. (2015). Computerized cognitive testing for older adults: a review. *American Journal of Alzheimer's Disease & Other Dementias*, *30*(1), 13-28.

FIGURES

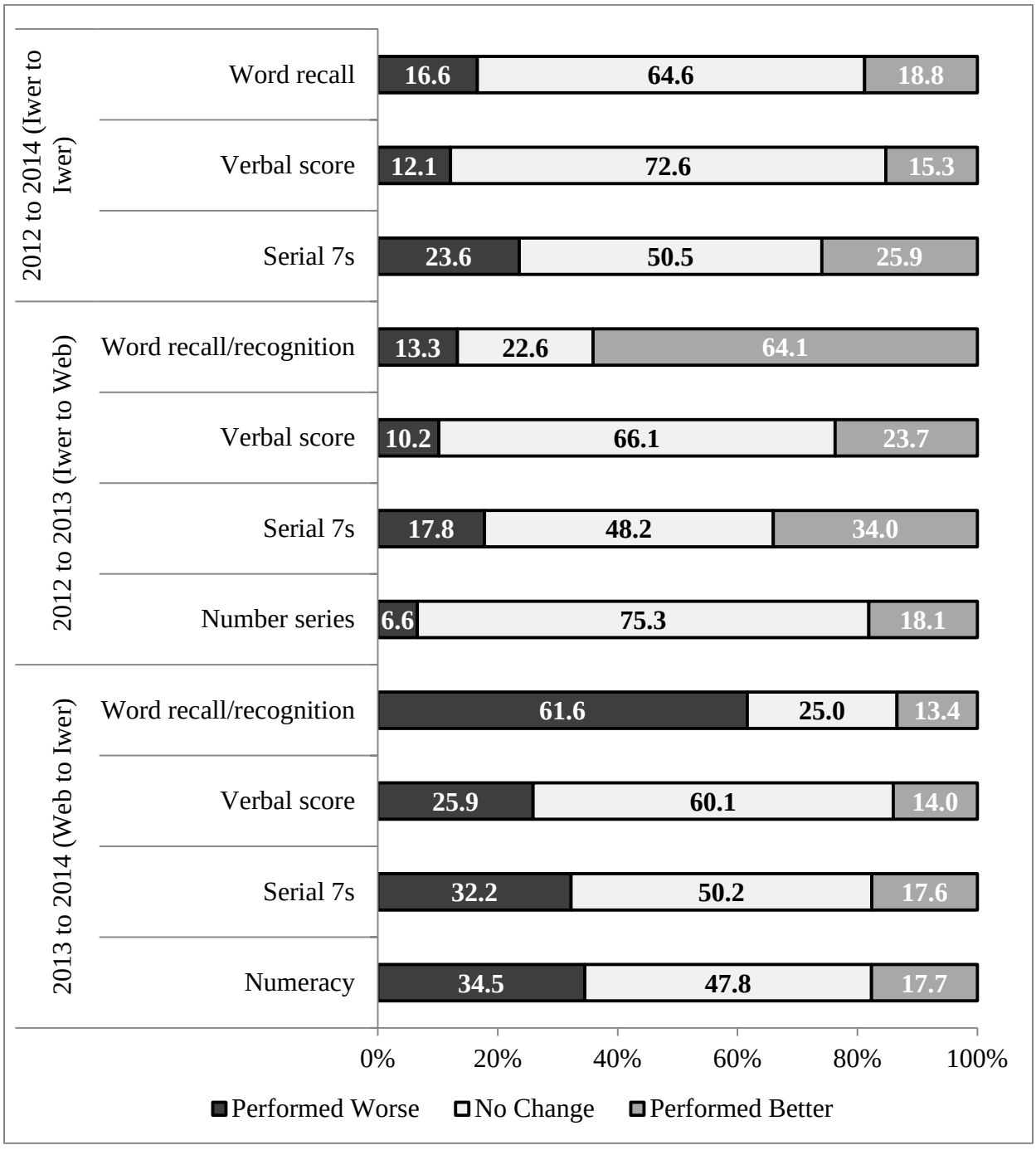

Figure 1. Percentage of respondents performing worse, better, and with no change between paired waves.

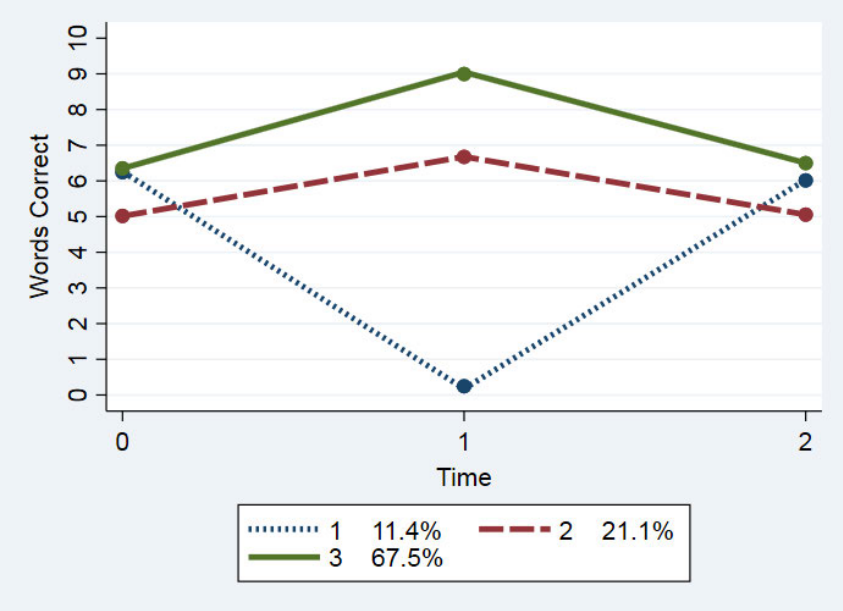

Figure 2. Plot of three-class solution, word recall/recognition

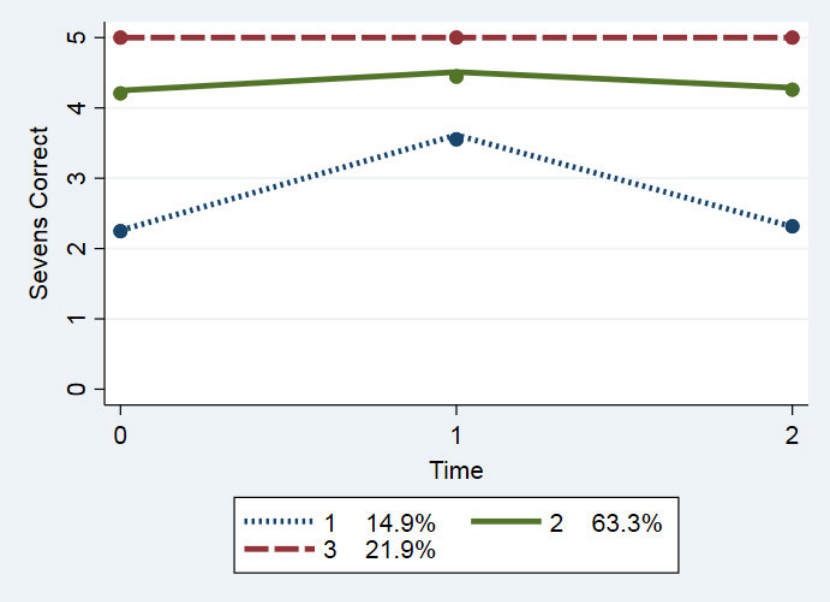

Figure 3. Plot of three-class solution, serial 7s

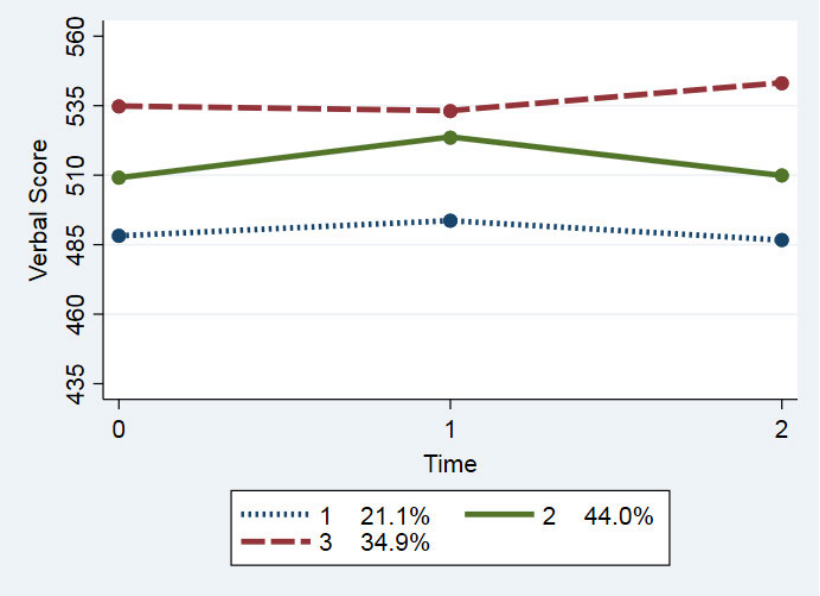

Figure 4. Plot of three-class solution, verbal analogies score

Table 1. Examples of cognitive tests involving self-administration in large-scale surveys

Study	Mode	Cognitive Tests
The Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS) (https://www.ncbi.nlm.nih.gov/ pubmed/24865195)	Self-administered (web) in group setting	Sensorimotor ability (motor praxis task), continuous performance and visual attention, reaction time (Emotional Stroop Test), impulse control (Go-No Go), emotion recognition, facial memory, attention and working memory (Short-Letter-N-Back Test), and abstraction and mental flexibility (conditional exclusion task)
Cognition and Aging In The USA (CogUSA) (http://www.icpsr.umich.edu/icpsrweb/ NACDA/studies/36053)	Mix of interviewer- and self-administered, multi-mode (face-to-face, telephone, web)	Telephone survey contains primarily HRS items. Face-to-face survey includes an extensive cognitive battery designed to measure attention, reaction time, processing speed, task switching, and inhibitory control (STOP and GO switching task); a vigilance task to measure attention and processing speed; and tests from the Woodcock Johnson Psychoeducational Test Battery (including retrieval fluency, verbal analogies, spatial relations, picture vocabulary, calculation, and concept formation) and the Wechsler Abbreviated Scale of Intelligence (vocabulary, block design, similarities, and matrix reasoning).
Health and Retirement Study (HRS) (http://hrs.isr.umich.edu/)	Interviewer- and self-administered, multi-mode (face-to-face and telephone for core waves; web for off-year studies)	Memory: Self-rated memory, immediate and delayed word recall; working memory: serial 7s, mental status: backwards count, date naming, object naming, person naming; abstract reasoning: similarities; fluid reasoning: number series; vocabulary; numeracy; retrieval fluency (animal naming); verbal analogies

Panel Study of Income Dynamics (PSID) (https://psidonline.isr.umich.edu/)	Interviewer- and self-administered, multi-mode (web 2016 supplemental study)	Quantitative number series, health and financial literacy (some items similar to numeracy scales), verbal reasoning (sentence completion)
UK BioBank (http://www.ukbiobank.ac.uk/)	Self-administered (touchscreen) within an interviewer-administered design	Prospective memory test; pairs matching test; fluid intelligence test; reaction time test; numeric memory test; lights pattern memory test; phonemic fluency/word category test; trail making test; symbol digit substitution test
Understanding America Study (https://uasdata.usc.edu)	Self-administered (web)	Self-rated memory, serial 7s, date naming, vocabulary, person naming, numeracy, number series, picture series, verbal analogies, risk preferences, decision making
Understanding Society (https://www.understandingsociety.ac.uk/)	Self-administered (CASI) within an interviewer-administered interview (primarily FTF); mix of CASI and web administration for innovation panel	Self-rated memory, episodic memory (immediate and delayed word recall), serial 7s, number series, verbal fluency, numeracy. Innovation panel only (wave 8): Prospective memory, phonemic fluency/word categories, serial 7s

Table 2. Sample size by mode for 2012, 2013, 2014 waves

Completed 2012 Core (50+, not via proxy)	+ Completed 2013 Web	+ Completed 2014 Core (not via proxy)	+Randomized to Mode (< 80, no mode switch)
FTF N: 10941	Web N: 2762	FTF N: 562 Tel N: 2072 Total N: 2634	Tel N: 2034
Tel N: 7791	Web N: 2558	FTF N: 2206 Tel N: 228 Total N: 2434	FTF N: 2189
Total N: 18372	Total N: 5320	Total N: 5068	Total N: 4223

Table 3. Characteristics of analytic sample compared to core respondents

	2012 Core Respondents (n=18372)	Analytic Sample (n=4223)
Female	58.3%	58.3%
Race/Ethnicity: Hispanic	13.3%	6.3%
Race/Ethnicity: NH Black	19.0%	12.0%
Race/Ethnicity: NH White, Other, NA	67.7%	81.7%
Education: Less than high school	20.4%	4.6%
Education: High school	31.4%	26.0%
Education: Some college	24.6%	30.3%
Education: College graduate or more	23.7%	39.1%
Working for pay	39.2% (n=18697)	54.7%
Chronic conditions: 0	29.3%	38.3%
Chronic conditions: 1	36.1%	37.6%
Chronic conditions: 2+	34.5% (n=18729)	24.2%
Age: Mean (SD), [Median]	67.0 (10.6) [Median: 65.0]	62.9 (7.6) [Median: 62.0]
Income: Mean (SD), [Median]	$65,363.6 ($101,526.0) [Median: $39557.7]	$96,758.3 ($124,830.1) [Median: $66,301.0]
Self-rated memory in 2012 (0-5, higher is better)	3.0 (1.0) (n=18714) [Median: 3.0]	2.8 (0.8) (n=4222) [Median: 3.0]

Cognitive Outcomes, 2012		
Quantitative number series	519.4 (33.0) (n=16208) [Median: 524.0]	535.2 (24.7) (n=4123) [Median: 536.0]
Serial 7s	3.4 (1.7) (n=18344) [Median: 4.0]	4.1 (1.3) (n=4190) [Median: 5.0]
Verbal analogies	501.0 (28.3) (n=1855) [Median: 498.0]	513.6 (24.9) (n=429) [Median: 516.0]
Word recall/recognition	5.3 (1.7) (n=18641) [Median: 5.0]	6.0 (1.5) (n=4213) [Median: 6.0]

Table 4. Analytic sample sizes for individual cognitive tests

Cognitive test	Analytic Sample Size
Tests administered in two waves to the same respondents:	
Quantitative number series (2012-2013 only)	994
Numeracy (2013-2014 only)	1069
Tests administered in all waves to the same respondents:	
Serial 7s	2147
Verbal analogies	429
Word recall/recognition	813[5]
Total eligible	4223

[5] In addition to random subsampling, some respondents screened out due to inability to hear the sound test necessary for administration, as described above.

Table 5. Percent of cases with item-missing data, and tests of differences

	2012		2013	2014	
	Tel % Missing (N)	**FTF % Missing (N)**	**Web % Missing (N)**	**Tel % Missing (N)**	**FTF % Missing (N)**
Number Series (First item) *(n=994)*	3.6% (19)	2.2% (10) *WEB*	5.3% (53) *F12*		
Numeracy 1 (Chance of disease) *(n=2030)*			0.6% (12) *T14,F14*	2.1% (21) *WEB*	1.8% (18) *WEB*
Numeracy 2 (Lottery split) *(n=2030)*			4.5% (91) *T14,F14*	7.2% (72) *WEB*	7.4% (76) *WEB*
Serial 7s (First item) *(n=2147)*	1.7% (19) *WEB*	0.8% (9)	0.6% (14) *T12*	1.1% (12)	1.4% (15)
Verbal Analogies (First item) *(n=429)*	0.0% (0) *WEB*	0.0% (0)	2.1% (9) *T12*	0.0% (0)	1.0% (2)
Word Rec. (Entire test) *(n=813)*	0.0% (0) *WEB*	0.1% (1) *WEB*	9.3% (76) *T12,F12, T14,F14*	0.0% (0) *WEB*	0.0% (0) *WEB*

Superscripts denote significant differences from difference-of-means tests ($p < .001$). Within-subject tests are conducted between web scores (WEB) and a given respondent's scores in telephone in 2012 (T12), FTF in 2012 (F12), telephone in 2014 (T14), and FTF in 2014 (F14). Between-subject tests are conducted comparing the scores of random subsets receiving FTF versus telephone administration within a wave.

Table 6. Completion times for cognitive tests (in seconds)

	2012		2013	2014		
Cognitive Test	**Tel Mean (SD)**	**FTF Mean (SD)**	**Web Mean (SD)**	**Tel Mean (SD)**	**FTF Mean (SD)**	**Mean Difference, Web - Iwer: 2013 - 2012, 2013 - 2014**
Number Series: *(n=970)*	366.93 *WEB* (106.51)	343.03 (102.70)	350.25 *T12* (178.22)			**- 5.60**
Numeracy Subset: *(n=1053)*			86.62 *T14,F14* (21.85)	68.64 *WEB* (18.01)	65.44 *WEB* (18.72)	**+ 19.60**
Serial 7s: *(n=2095)*	38.13 *WEB* (14.69)	35.27 *WEB* (14.13)	69.02 (35.28) *T12,F12,T14,F14*	37.78 *WEB* (13.78)	34.88 *WEB* (14.82)	**+ 32.24, + 32.72**
Verbal Analogies: *(n=422)*	97.13 *WEB* (31.42)	89.76 *WEB* (29.75)	148.56 (77.34) *T12,F12,T14,F14*	82.01 *WEB* (25.34)	76.18 *WEB* (26.73)	**+ 55.27, + 69.10**
Word Rec.: *(n=757)*	79.06 *WEB* (13.65)	75.40 *WEB* (13.28)	209.08 (81.65) *T12,F12,T14,F14*	80.16 *WEB* (13.10)	75.71 *WEB* (13.00)	**+ 131.2, + 131.8**

Sample size is restricted to all respondents shown a given test in all three waves with a positive time spent in each wave. Comparisons against web administration are within-subject tests; comparisons between TEL and FTF administration are between-subject tests. Subscripts refer to significant differences from difference-of-means tests (all significant at p<.001).

Table 7. Means, standard errors, and tests of differences for cognitive scores

	2012		2013	2014	
	Tel Mean (SE) *(N)*	**FTF Mean (SE) *(N)***	**Web Mean (SE) *(N)***	**Tel Mean (SE) *(N)***	**FTF Mean (SE) *(N)***
Number Series	535.03 (1.08) *(520)* ^{WEB}	532.50 (1.24) *(453)* ^{WEB}	541.38 (0.71) *(973)* ^{T12,F12}		
Numeracy			2.95 (0.03) *(1069)* ^{T14,F14}	2.56 (0.05) *(526)* ^{WEB}	2.67 (0.05) *(543)* ^{WEB}
Serial 7s	4.12 (0.04) *(1072)* ^{WEB}	4.05 (0.04) *(1041)* ^{WEB}	4.43 (0.02) *(2113)* ^{T12,F12,T14,F14}	4.20 (0.04) *(1041)* ^{WEB}	4.07 (0.04) *(1072)* ^{WEB}
Verbal Analogies	512.00 (1.68) *(201)* ^{WEB}	515.18 (1.74) *(212)* ^{WEB}	520.52 (1.18) *(413)* ^{T12,F12,T14,F14}	513.87 (1.92) *(212)* $^{WEB, F14}$	519.43 (1.90) *(201)* ^{WEB,T14}
Word Rec.	6.11 (0.07) *(421)* ^{WEB}	5.99 (0.07) *(392)* ^{WEB}	7.50 (0.10) *(813)* ^{T12,F12,T14,F14}	6.10 (0.08) *(392)* $^{WEB, F14}$	6.17 (0.07) *(421)* $^{WEB, T14}$

Superscripts are defined in previous table.

Table 8. Proportion achieving maximum score for selected tests

	2012		2013	2014	
	Tel % (N)	**FTF % (N)**	**Web % (N)**	**Tel % (N)**	**FTF % (N)**
Numeracy			41.2% (441)	24.7% (130)	30.0% (163)
Serial 7s	57.6% (617)	53.8% (560)	68.4% (1446)	58.4% (608)	53.9% (1072)
Word recall/recognition	1.2% (5)	0.3% (1)	26.6% (216)	2.0% (8)	1.0% (4)

Table 9. Within-test correlations across waves of data collection

	2012		2014		2012/2014
Cognitive Test	**Tel*2013 Web**	**FTF*2013 Web**	**Tel *2013 Web**	**FTF *2013 Web**	**Iwer*Iwer**
Number Series	0.42 (520)	0.51 (453)			
Numeracy			0.49 (526)	0.46 (543)	
Serial 7s	0.22 (1072)	0.27 (1041)	0.31 (1041)	0.22 (1072)	0.52 (2113)
Verbal Analogies	0.43 (201)	0.41 (212)	0.44 (212)	0.33 (201)	0.63 (413)
Word Rec.	0.06 (421)	0.12 (392)	0.12 (392)	0.24 (421)	0.31 (813)

Sample size restricted to respondents who saw and substantively answered a given test in all waves of administration.

Table 10. Between-test correlations within waves of data collection

	Word rec. ρ (N)	Serial 7s ρ (N)	Analogies ρ (N)	Num. series ρ (N)	Rate memory ρ (N)
Correlations with self-rated memory					
2012 Tel	0.21 (420)	0.10 (1072)	0.12 (200)	0.12 (518)	
2012 FTF	0.10 (391)	0.13 (1038)	0.28 (212)	0.13 (453)	
2013 Web	0.05 (811)	0.08 (2110)	0.18 (412)	0.12 (992)	
2014 Tel	0.16 (391)	0.16 (1038)	0.14 (212)		
2014 FTF	0.20 (420)	0.15 (1072)	0.12 (200)		
Correlations with word recall or recognition					
2012 Tel			0.21 (75)	0.20 (204)	0.21 (420)
2012 FTF			0.30 (68)	0.24 (192)	0.10 (391)
2013 Web			0.21 (143)	0.27 (396)	0.05 (811)
2014 Tel			0.13 (68)		0.16 (391)
2014 FTF			0.22 (75)		0.20 (420)
Correlations with numeracy					
2013 Web		0.27 (1066)			0.14 (1067)
2014 Tel		0.40 (526)			0.13 (524)
2014 FTF		0.51 (540)			0.18 (543)

Sample size restricted to respondents who saw and substantively answered both relevant tests in all waves of administration.

Table 11. Longitudinal models for cognitive outcomes

	Word Recall/ Recognition	Serial 7s	Verbal Score
Fixed Effects			
Intercept	6.045 (0.088)***	4.081 (0.032)***	513.73 (1.624)***
Time: 2013 (vs. 2012)	1.445 (0.102)***	0.345 (0.032)***	6.889 (1.335)***
Time: 2014 (vs. 2012)	0.080 (0.094)	0.043 (0.028)	2.947 (1.255)*
2012 FTF (vs. telephone)	0.030 (0.099)	0.017 (0.038)	-0.183 (2.018)
Variance Components			
Random Intercept	0.188	0.252	272.64
Autoregressive Errors	0.175	0.264	0.1302
Residual Variance	4.328	1.137	374.08
Model Fit & Sample Size			
BIC	10553.6	19367.4	11279.6
N	813	2113	413

***p<.001, **p<.01, *p<.05

Table 12. Predictors of cognitive scores (continued on next page)

	Number 12	Number 13	Numeracy 13	Numeracy 14	Sevens 12	Sevens 13	Sevens 14
Intercept	567.717	530.427	-2.259	-0.292	1.313	0.656	0.525
	(51.018)***	(56.641)***	(2.233)	(2.542)	(1.946)	(1.606)	(1.926)
Age (continuous)	-1.341	0.011	0.142	0.051	0.059	0.105	0.079
	(1.592)	(1.768)	(0.071)*	(0.08)	(0.062)	(0.051)*	(0.061)
Age squared (continuous)	0.01	-0.003	-0.001	0.000	0.000	-0.001	-0.001
	(0.012)	(0.014)	(0.001)*	(0.001)	(0.000)	(0.000)*	(0.000)
Female (vs. Male)	-5.509	-5.192	-0.301	-0.203	-0.19	-0.044	-0.079
	(1.4)***	(1.558)***	(0.061)***	(0.07)**	(0.054)***	(0.045)	(0.054)
Race/Ethnicity: Hispanic (vs. White/Other/NA)	-11.526	-12.489	-0.612	-0.593	-0.699	-0.038	-0.563
	(2.887)***	(3.205)***	(0.132)***	(0.151)***	(0.11)***	(0.09)	(0.109)***
Race/Ethnicity: Non-Hispanic Black (vs. White/Other/NA)	-10.638	-16.691	-0.719	-0.588	-0.72	-0.338	-0.619
	(2.249)***	(2.505)***	(0.112)***	(0.127)***	(0.084)***	(0.069)***	(0.083)***
Education: College grad + (vs. Less than high school)	14.481	24.714	0.603	0.972	0.519	0.302	0.448
	(3.518)***	(3.92)***	(0.177)***	(0.202)***	(0.132)***	(0.108)**	(0.13)***
Education: High school (vs. Less than high school)	4.880	9.435	0.193	0.456	0.085	0.191	0.06
	(3.529)	(3.926)*	(0.181)	(0.206)*	(0.131)	(0.108)+	(0.129)
Education: Some college (vs. Less than high school)	11.024	15.66	0.431	0.706	0.368	0.148	0.285
	(3.492)**	(3.89)***	(0.178)*	(0.203)***	(0.131)**	(0.108)	(0.129)*
Income: Fourth quartile (vs. first)	3.597	3.585	0.143	0.26	0.189	0.009	0.215
	(2.048)+	(2.289)	(0.096)	(0.114)*	(0.085)*	(0.069)	(0.084)*
Income: Second quartile (vs. first)	1.229	1.56	0.001	-0.004	0.096	0.054	0.113
	(1.913)	(2.127)	(0.095)	(0.107)	(0.078)	(0.064)	(0.076)
Income: Third quartile (vs. first)	1.303	1.917	0.024	0.008	0.043	-0.016	0.078
	(2.066)	(2.298)	(0.092)	(0.109)	(0.078)	(0.064)	(0.079)
Currently working for pay (vs. not)	-1.194	0.038	0.04	-0.016	0.007	0.017	0.102
	(1.577)	(1.784)	(0.072)	(0.081)	(0.062)	(0.052)	(0.062)+
CESD score (continuous)	-1.1	-1.228	-0.046	-0.044	-0.06	-0.057	-0.046
	(0.412)**	(0.457)**	(0.019)*	(0.021)*	(0.016)***	(0.013)***	(0.015)**
Chronic conditions, set of four: One condition (vs. none)	-2.784	-2.387	0.051	0.009	0.079	0.027	0.063
	(1.561)+	(1.732)	(0.068)	(0.079)	(0.061)	(0.05)	(0.061)
Chronic conditions, set of four: Two or more conditions (vs. none)	-1.923	-5.058	-0.059	-0.075	-0.046	-0.025	0.014
	(1.842)	(2.045)*	(0.084)	(0.092)	(0.072)	(0.06)	(0.069)
Count of internet activities engaged in	2.7	1.109	0.108	0.132	0.104	0.077	0.095
	(0.559)***	(0.618)+	(0.028)***	(0.031)***	(0.023)***	(0.019)***	(0.023)***
N	973	973	1069	1069	2113	2113	2113
R squared	0.218	0.170	0.185	0.168	0.137	0.054	0.113

Table 12. Predictors of cognitive scores (continued)

	Verbal 12	Verbal 13	Verbal 14	Word 12	Word 13	Word 14
Intercept	545.009	619.35	502.487	0.34	-1.103	-8.176
	(89.07)***	(84.821)***	(101.095)***	(3.813)	(7.808)	(3.933)*
Age (continuous)	-1.698	-3.329	-0.309	0.16	0.279	0.444
	(2.793)	(2.66)	(3.162)	(0.12)	(0.246)	(0.124)***
Age squared (continuous)	0.012	0.022	-0.001	-0.001	-0.002	-0.004
	(0.022)	(0.021)	(0.024)	(0.001)	(0.002)	(0.001)***
Female (vs. Male)	3.239	4.65	3.952	0.484	0.085	0.504
	(2.317)	(2.207)*	(2.618)	(0.104)***	(0.213)	(0.107)***
Race/Ethnicity: Hispanic (vs. White/Other/NA)	-15.284	-11.813	-9.84	-0.272	-0.372	-0.061
	(4.451)***	(4.236)**	(5.019)+	(0.204)	(0.418)	(0.209)
Race/Ethnicity: Non-Hispanic Black (vs. White/Other/NA)	-14.46	-16.693	-16.696	-0.496	-1.454	-0.237
	(3.49)***	(3.307)***	(3.978)***	(0.174)**	(0.356)***	(0.178)
Education: College grad + (vs. Less than high school)	26.669	19.661	25.382	1.01	0.34	0.683
	(5.719)***	(5.442)***	(6.439)***	(0.268)***	(0.548)	(0.273)*
Education: High school (vs. Less than high school)	5.231	2.585	9.239	0.436	-0.525	0.03
	(5.663)	(5.386)	(6.367)	(0.269)	(0.549)	(0.274)
Education: Some college (vs. Less than high school)	14.83	6.843	15.015	0.643	-0.063	0.34
	(5.653)**	(5.374)	(6.357)*	(0.268)*	(0.547)	(0.272)
Income: Fourth quartile (vs. first)	-6.837	-0.576	1.678	0.066	0.166	0.541
	(3.63)+	(3.428)	(3.941)	(0.161)	(0.323)	(0.163)***
Income: Second quartile (vs. first)	-3.781	-1.431	-1.914	-0.022	0.241	0.2
	(3.195)	(3.039)	(3.632)	(0.15)	(0.305)	(0.16)
Income: Third quartile (vs. first)	-4.875	-1.743	0.092	0.053	0.266	0.36
	(3.269)	(3.109)	(3.689)	(0.151)	(0.306)	(0.15)*
Currently working for pay (vs. not)	-0.12	-4.863	4.604	0.209	0.231	-0.085
	(2.754)	(2.547)+	(3.076)	(0.126)+	(0.245)	(0.125)
CESD score (continuous)	-1.344	-2.607	-0.528	-0.016	-0.139	-0.082
	(0.71)+	(0.674)***	(0.714)	(0.033)	(0.068)*	(0.033)*
Chronic conditions, set of four: One condition (vs. none)	0.028	0.173	0.661	0.054	-0.299	-0.03
	(2.688)	(2.561)	(3.125)	(0.119)	(0.244)	(0.124)
Chronic conditions, set of four: Two or more conditions (vs. none)	-0.591	1.113	-0.897	-0.076	-0.189	-0.118
	(3.011)	(2.87)	(3.365)	(0.141)	(0.29)	(0.143)
Count of internet activities engaged in	3.686	3.798	3.996	0.011	0.197	0.006
	(1.044)***	(0.99)***	(1.158)***	(0.046)	(0.094)*	(0.047)
N	413	413	413	813	813	813
R squared	0.244	0.272	0.238	0.100	0.085	0.123