Studying Cognitive Aging and Dementia in the Health and Retirement Study: Design and Measurement Strategies to Improve Research

M. Maria Glymour[1] and Scott Hofer[2]

1. Department of Epidemiology and Biostatistics
University of California, San Francisco
San Francisco, CA 94158
mglymour@epi.ucsf.edu

2. Department of Psychology
University of Victoria
Victoria, BC
Canada
smhofer@uvic.ca

May 17, 2016

Summary

The Health and Retirement Study (HRS) is a unique resource for cognitive aging research due to the diverse and representative sample; the detailed social and physical health measures; and the long-term follow-up. Applications of HRS data in cognitive aging research have been widespread, but the brief cognitive assessments in the core battery represent an important limitation. These assessments are not sensitive to change in high-functioning individuals, scores improve with repeated test encounters, and assessments have substantial measurement error (reduces statistical power). Although it would be valuable to enhance the existing data to allow for higher quality dementia diagnoses, this addition would not solve the most important challenges for cognitive aging research in HRS. The most useful contemporary analytic approaches to identify factors that influence dementia risk are based on assessing within-person change in cognitive function (measured as a continuous variable). These efforts must begin follow-up years before an individual meets diagnostic criteria. Adding the expanded Harmonized Cognitive Assessment Protocol (HCAP) to the HRS battery will be invaluable, and assessments in racial/ethnic minorities will be of special import for the field. Among several possible options to implement the HCAP, a measurement burst design -- with multiple closely spaced repeat assessments at each wave -- would address potential retest bias. Such a design may not be logistically feasible, however, so we also discuss other design options.

Given the cost tradeoffs entailed by the additional interview burden, a mixed/adaptive design for the HCAP may be optimal. An adaptive design would have two key features: randomly staggered introduction of cognitive assessments to allow for population estimates of retest effects; and more frequent assessments for individuals likely to be declining quickly. For example, healthy participants, for whom little annual cognitive change is expected, could receive the full HCAP every 4 years. The fraction of participants with evidence of incipient cognitive decline might be interviewed much more frequently, depending on resources, every 6 months or annually. Two instruments from the HCAP battery, the animal naming and digit span assessments, should be considered as candidates for the core HRS battery fielded every two years. These two brief elements provide a good tradeoff of new information for minimal additional interview time.

Additional potential enhancements to HRS, summarized in Table 1 on page 11, roughly in order of least to most challenging, include: release additional interview meta-data that predict cognitive scores; standardize imputation using information from external sources on cognitive status of participants who attrite, e.g., dementia diagnosis on death records or Medicare; add a proxy interview for a sample of individuals who also participate in direct cognitive assessments; enhance literacy/premorbid cognition measures in the core interview; link to external data with information on premorbid cognition; add neuroimaging or neuropathology on selected sample members; and field a web-based or passive cognitive assessment strategy. Some of these possibilities might not be feasible due to expense or participant burden, but other strategies are relatively inexpensive opportunities to enhance the value of HRS for research on the determinants, prognosis, and consequences of cognitive changes in middle-aged and older adults.

Introduction

We were asked to consider the core cognitive assessments in the Health and Retirement Study (HRS) and make recommendations for future priorities. Based on the questions asked of us, we organized our responses into four topics:

1. Potential Contributions of HRS for Cognitive Aging Research: What is the purpose of the HRS cognitive measures? How could they be used now and in the future? What unique features of HRS could be used to fill evidence gaps impossible to address with other cohorts?
2. Limitations of Current Approaches: What are limitations of current HRS design and measures? Are there specific research areas which are currently impossible or severely limited by the cognitive assessments in HRS?
3. Optimal Timing for HCAP: Given the range of cognition measures collected in the HRS core content, what is the spectrum of options for how frequently the HRS should conduct the Harmonized Cognitive Assessment Protocol (HCAP)?
4. Enhancing Core Content to Leverage HCAP: What is missing from the HRS core content to allow measurement of dementia and cognitive aging in between HCAP assessments, for example informant interviews on respondents with probable mild cognitive impairment? We consider both relatively simple (with respect to time and expense) changes and potentially high impact but more expensive/challenging changes to HRS core cognitive assessments.

1. Potential Contributions of HRS for Cognitive Aging Research

HRS plays a unique role in the US research environment and since its launch has been used in diverse research areas. The cognitive data are no exception. HRS cognitive measures are relevant for research on several heterogeneous categories of research questions:

- Identifying risk or protective factors (including demographic, socioeconomic, behavioral, and medical variables, among others) for individual change with the goal of identifying opportunities to prevent or delay cognitive aging, including Alzheimer's disease and associated disorders (ADAD).
- Predictors of prognosis for physical and cognitive functioning after cognitive losses or onset of ADAD.
- Research on formal and informal care for people with cognitive loss or ADAD, including caregiver burden.
- The impact of cognition on social, financial, or other health outcomes.
- Descriptive studies of neuropsychological processes or cognitive aging.
- Development or validation of screening/predictive models.

As noted in previous work, the original goals for the HRS cognitive assessments were to describe cognition in a population sample of middle aged and older adults; measure functioning

across a range from severe impairment to healthy cognition; and reflect changes over time in cognitive function. In addition, because of the HRS design, it was important that the measures could be administered over the telephone with a limited time allocation[1, 2]. The relative importance of careful measurement of cognition has grown since the inception of HRS. This reflects the increasing recognition of the prevalence and burden of cognitive impairment in older adults, and research showing that cognitive function correlates with a host of other social and physical measures of well-being, such as financial security[3, 4]. Thus, it is reasonable that the resources devoted to careful cognitive assessments have also grown.

The addition of the ADAMS sample allowed us to confirm that the HRS core cognitive measures provide reasonably good screeners for dementia[5], especially when incorporating the proxy respondent information[6]. This addition served to improve the credibility of HRS-based research to the neurology and neuropsychology research communities and facilitated calibration against many other studies.

HRS is not primarily a cognitive aging study, but it has several advantages over most other studies that include cognitive measures. Because it is nationally representative, it includes more racial, ethnic, and geographic diversity than most studies. For example, HRS includes a large contingent of Latinos. HRS may also have a large number of Asian-Americans compared to other US cohorts, but this information is not available in the public use data. The geographic diversity has facilitated novel studies that could not plausibly be conducted in other samples[8] and has provided important complementary information when analyzed in parallel with other cohorts[9]. The longitudinal design, with nearly 25 years of continuous follow-up on the first enrollment cohort, makes a unique resource for description of long-term changes. HRS has a sufficiently large accumulated sample that even relatively rare events, such as stroke, can be studied in relation to cognition[10]. HRS includes much more comprehensive socioeconomic and social measures than most studies. The flexible and dynamic use of measurement modules to pilot and validate new measures has substantially enhanced the value of the data.

There is growing interest in passive or administrative data sources, such as electronic health records, but active-data collection cohorts such as HRS have important strengths over passive data collection sources. Most importantly, social risk factors and outcomes are generally much better measured in HRS. Further, many critical health outcomes, including disability, depressive symptoms, and (for our purposes most important) cognition, are rarely available in passive-data collection settings. Although participation in cohorts such as HRS is almost certainly influenced by various health risk factors, the selection process is likely quite distinct from the processes that influence participation in passive data collection sources, e.g., Kaiser, so there is a huge advantage in comparing evidence from both types of data. The potential impact of research combining evidence from passive/administrative data sources with evidence from active cohorts such as HRS is important when considering how to provide the most powerful data with HRS. HRS will often be used in analyses intended to complement work in other data sources, for example to rule out sources of bias in other data or to confirm generalizability of findings from cohorts with narrower or non-representative sampling frames. Because of this unique advantage of HRS, including measures that facilitate calibration and

harmonization with other cohorts is important. In other words, even measures that are obviously not the best available instruments may be valuable if they allow us to map performance of people from another data source onto the demographic distribution in HRS.

2. Limitations of Current Approaches

A major emphasis of research on aging-related processes concerns understanding how and why psychological, physiological and behavioral phenomena change over time within and between aging individuals. This goal is complicated by the possibility that observable change in any given individual may reflect the joint influences of multiple processes and also be affected by differences in study design, measurements, and sample characteristics[11]. For example, observable and quantifiable decreases in memory performance over time (i.e., with increasing age) may reflect the additive or synergistic effects of declining vascular health and early stage progression of Alzheimer's dementia, but may also be attenuated by performance gains attributable to repeated testing. In regards to the HRS, the current core content has performed extraordinarily well given the logistical constraints of a brief, telephone-based, assessment. The major limitations, some of which are common to other cohorts that emphasize the study of dementia and cognitive impairment are:

- Severe ceiling effects on the Telephone Interview for Cognitive Status (TICS) make it difficult to detect early cognitive changes. This upper scoring boundary (ceiling) is somewhat obscured when all cognitive measures are summed, as is common practice, but when using such a sum, most of the variance is due to word list recall.
- Substantial retest effects can make it difficult to detect decline and may bias estimated effects of risk factors on rate of change (if those risk factors are also related to retest effects).
- Measurement error in brief cognitive assessments results in lower power to detect within-person change. Although the larger sample size of the HRS offsets this limitation for population estimates of slope variance[12] (see Rast and Hofer Table 6), a substantial power problem remains when trying to identify determinants of rates of change (i.e., compare slopes between groups).
- Selective attrition, due to cognitive impairment or illness that preclude direct assessments, study dropout, and mortality bias longitudinal analyses. These three sources of attrition all introduce bias, but the range of methodological responses to alleviate bias may be different[11]. Because cognition is such a strong predictor of dropout and mortality, this type of bias can be important.
- Limited measurement of premorbid cognitive function, even of constructs that would strongly predict old age cognitive function (e.g., literacy). This also presents an important challenge for assessments of cognitive change because much of the between-person variance in cognitive scores is due to pre-morbid functioning or test-taking skills[13].
- The full cognitive assessment is only performed once before age 65, so any change before age 65 is not detectable. While there are limitations for sensitive detection of change in higher functioning individuals, if more reliable and sensitive measures were added to the core battery, it would be useful to incorporate longitudinal assessments

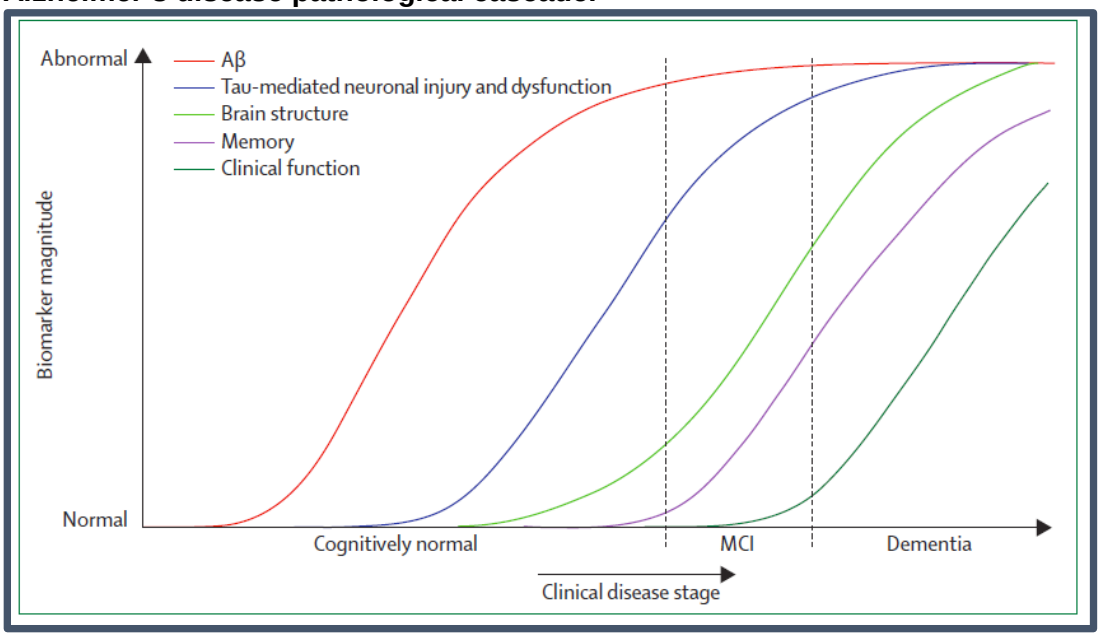
from enrollment on. This would enable more sensitive detection of changes in midlife potentially related to premorbid dementia and other health issues.

- Uncertain performance of the brief cognitive measures across sociodemographic groups defined by race/ethnicity, education, or adult SES[14, 15]. For example, these measures may be more specific for organic brain disease among whites than blacks. ADAMS did not include enough Latinos to allow for estimation models and evidence suggested that the correspondence between core HRS assessments and diagnosis of dementia differed for Latinos versus non-Latinos[6].

ADAMS included a categorization of dementia type, including probable AD, possible AD, and vascular dementia. Although ADAMS used state of the art methods at the time, those methods nonetheless reflect the ongoing evolution in our understanding of the contributions of vascular and Alzheimer's processes to dementia. Even with comprehensive neuropsychological examinations, it is difficult to distinguish these processes and many (perhaps the vast majority of) older people with cognitive losses have been affected by both vascular and AD-related processes. In the absence of neuroimaging or neuropathological measures, there are intrinsic limitations to the capacity to evaluate underlying brain disease based on clinical symptoms (with current evidence, these limitations are not even fully resolved by neuroimaging or neuropathology).

Further, in the years since the launch of HRS, our conceptualization of dementia and in particular AD has changed substantially. Clinical AD is now recognized to be the late manifestation of a long-developing process, which originates years or even decades prior to diagnosis. This is represented in the highly cited Jack model of dynamic biomarkers for AD (reproduced in figure 1)[16]. Although this model has undergone revision since original publication, the central point of relevance for HRS design is the long period of accumulating pathophysiology and deteriorating memory that precedes the diagnosis of AD or even Mild Cognitive Impairment (MCI). The time scale is still an active research area, but changes probably typically begin decades before diagnosis[17]. As much as 12 years prior to diagnosis, verbal fluency measures decline at an accelerated rate in people with prodromal dementia compared to age-peers[18]. This implies that research using a dementia screen as the outcome is focused too late to identify causes of dementia. Indeed, nearly any binary outcome assessment will reflect a somewhat arbitrary cutpoint in a cumulative, continuous degenerative process. For this reason, there is increasing emphasis on assessing rates of change in cognition or other biomarkers, rather than incidence of dementia or other diagnostic outcome.

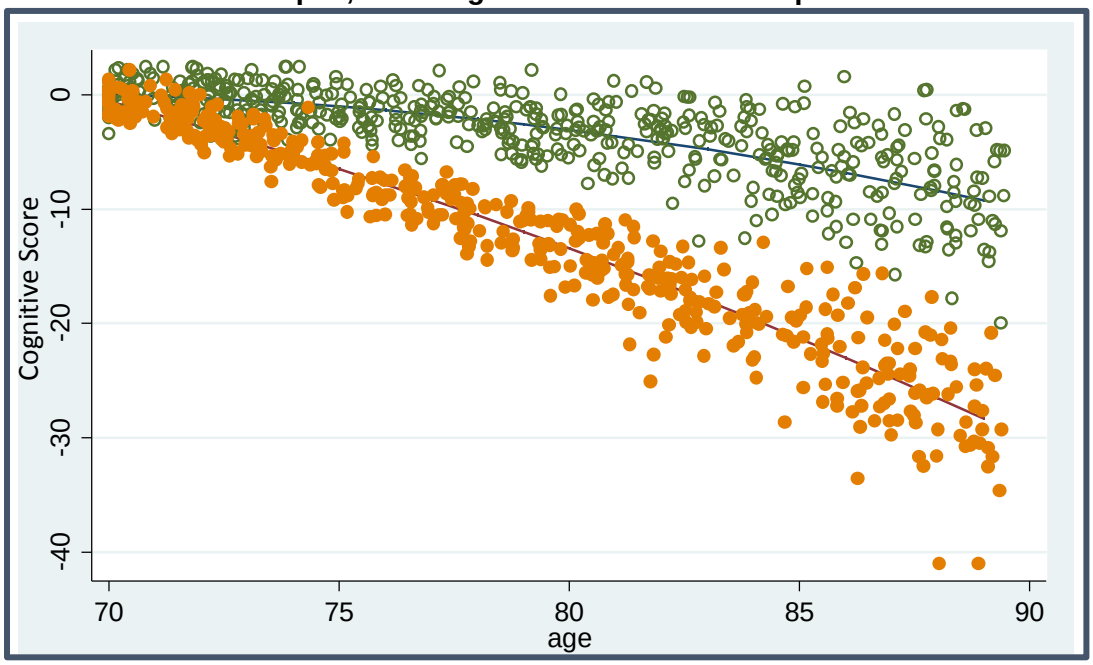
Figure 1. Reproduced from Jack, 2013: 2010 model of dynamic biomarkers of the Alzheimer's disease pathological cascade.



In the context of aging and neurodegenerative disorders, it is essential to differentiate change and impairment relative to an individual's earlier typical or maximal levels of functioning. The limitations of norm-referenced assessments are particularly significant when identifying subtle changes in heterogeneous pre-clinical populations, particularly in high-functioning individuals. Because AD and other types of dementia develop over years or decades, with neuropsychological consequences well before individuals meet diagnostic criteria, research on the causes of dementia is most robust when using within-person change in cognitive test scores as outcomes. Perhaps counterintuitively, dementia or AD diagnoses are not necessarily the optimal outcomes for most research on the causes of dementia or AD. The age at which an individual meets diagnostic criteria is strongly influenced by processes unrelated to the underlying pathology driving cognitive deterioration, so onset of impairment or dementia is not as informative as within-person change.

Because evaluating within-person change is so important, it is critical that the noise or random error in cognitive assessments at each wave be minimized. This requires a combination of sensitive measures and more frequent observations than is typically obtained in longitudinal studies of aging and dementia.The primary interest in research on determinants of cognitive change is difference in slopes across time, so the critical determinants of ideal design are factors that influence the variance of the slope estimates. These factors include: measurement error at each cognitive assessment; the steepness of slopes; the time elapsed between successive assessments; and sample size[12](figure 2). The steeper the average slope, the easier it will be to detect a slope difference of any absolute magnitude. More frequent measures add little advantage for people who are changing slowly (assuming those respondents do not drop out or die between consecutive measures).

Figure 2. Typical analyses aim to compare slopes in cognitive function for individuals with versus without a hypothetical exposure (represented by orange versus green dots). Power is better when there is less measurement error/noise in the assessments, greater differences in the slopes, and longer duration of follow-up.



3. Optimal Timing for the Harmonized Cognitive Assessment Protocol (HCAP):

Given the range of cognition measures collected in the HRS core content, what is the spectrum of options for how frequently the HRS should conduct the HCAP? Cognitive impairments constitute an increasing objective and subjective problem with advancing age. Laske (2012) in NEJM notes "what is still missing is the identification of a non-invasive, inexpensive diagnostic tool that could be used for population-wide screening to identify persons with preclinical Alzheimer's disease who are still cognitively healthy"[19]. While use of norm-referenced assessment tools that compare an individual against an average or typical person are useful in measuring cognitive status, the most powerful and sensitive approach for identifying individual change is repeat measurement of cognitive functioning. However, for identification of individual change, frequent assessments are necessary to reliably identify acceleration in changes (i.e., change-points) in cognitive functioning from an individual's prior established level or prior rate of change.

Several advantages of more frequent assessments:

- More statistical power to detect cognitive change
- More likely to detect earliest changes, and thus to identify factors that are causes of disease process rather than consequences of disease
- Capacity to model selection processes due to incapacity or mortality
- Usable research data accumulate more quickly

must be weighed against the disadvantages of more frequent assessments:

- Cost
- Participant burden
- Potential exacerbation of retest effects (uncertain)
- For a given number of assessments, closely spaced time intervals increase the uncertainty in slope estimates, i.e., if you only field 3 assessments, you learn more if they are fielded in years 0, 2, and 4 than years 0, 1 and 2.

Long-term widely spaced longitudinal designs conflate retest effects and long-term change[20]. Measurement burst designs [21-24], which feature sets of measurements comprising a number of closely spaced assessments, address this problem. In such a design, one burst might consist of multiple assessments conducted on a daily or weekly basis. The bursts themselves are spaced over longer intervals such as, for example, months or years. This design allows one to effectively separate short-term (e.g., day-to-day) within-person variability from long-term (year-to-year) within-person level, change, and variation. The strength of the burst design is the ability to improve the precision of the estimate by measuring a variable repeatedly over a short period of time. By doing so, the power for the detection of long-term change in burst studies can be increased, and they can be designed with shorter intervals and fewer subjects compared to multiwave designs.

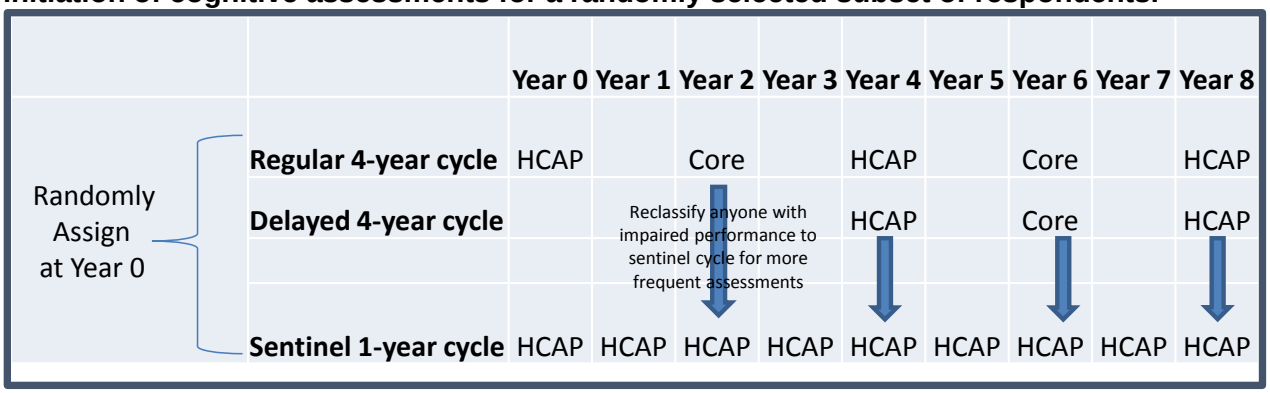
Measurement burst designs may prove too expensive and burdensome for HRS participants however. An alternative option adopts a mixed or adaptive model for the HCAP follow-up frequency. The ability to tailor assessment intervals to individuals at risk provides an alternative "adaptive longitudinal" design to best capture individual change and reduce burden on the overall sample.

As an example of such a design, at baseline individuals could be randomly assigned to 1 of 3 groups:

1. Regular 4-year cycle: HCAP every 4 years (or even every 6 years based on cost considerations), with assessments initiated as soon as possible after enrollment.
2. Delayed start 4-year cycle: HCAP every 4 years, with delayed initiation by 1 cycle to allow for evaluation of retest effects.
3. High frequency sentinel subset: More frequent assessments as often as affordable, perhaps every 6 months or annually.

The majority of individuals would likely be assigned to the regular 4-year cycle, with a smaller group on the delayed cycle and an even smaller group in the sentinel category with frequent follow-ups. For each regular or delayed cycle individual, predict at every core or HCAP assessment his or her composite cognitive score based on education, age and previous cognitive measures (including prior slope). Anyone who scores substantially (e.g., more than 1 standard error) below their predicted value should be recategorized into the high-frequency sentinel group (assuming consent) for more frequent follow-ups.

Figure 3. Example of a mixed/adaptive design alternating HCAP and core cognitive assessments for HRS participants. This design includes alternating HCAP and core assessments, with two novel features: (1) a more frequent "sentinel" assessment cycle for a randomly selected subsample of participants, with reclassification of participants experiencing rapid declines to the high-frequency assessment cycle, and a (2) delayed initiation of cognitive assessments for a randomly selected subset of respondents.



This model reduces participant burden because the full HCAP protocols will be rare for most respondents. It allows for evaluation of retest effects as a population level. The sentinel subset allows us to model selection processes, e.g., dropout of most cognitively impaired or mortality predicted by cognitive decline, with some detail. These estimates can be incorporated into models to correct for known biases. Converting follow-up to the sentinel subset when there is evidence of impairment will allow us to evaluate processes such as terminal decline. A slightly more complex mixed model might have more frequent assessments for older adults, even if they were performing well. Every 4 years is too long an interval for 90 year olds (roughly half of whom will die in a 4-year interval), although fine for most 50 year olds.

The major disadvantage of a mixed approach is the logistical complexity of implementation, and the potential for users to incorrectly analyze the data (the design implies there will be more and better assessments for impaired individuals). The researchers who manage HRS are of course excellent at guiding users on wise analyses to account for design effects, but this may merit special attention, e.g., considering the data from the reclassified sentinel subjects as sensitive data (in the same category as the biomarkers are currently treated).

The above design, including one group randomly assigned to a delayed first encounter with the cognitive battery, is envisioned for new enrollees. For current HRS participants, the delayed cycle is not useful because it is impossible to avoid retest effects. A mixed design with different assessment frequencies is relevant for current enrollees as for new enrollees. That is, existing HRS participants, if cognitively healthy, could be randomly assigned to either a regular 4-year cycle or a more frequent sentinel cycle. Current HRS participants whose past cognitive scores suggest they are entering a stage of rapid deterioration would be assigned to the high-frequency sentinel group.

4. Enhancing Core Content to Leverage HCAP

We were asked to consider what is missing from the HRS core content to allow measurement of dementia and cognitive aging in between HCAP assessments, for example informant interviews on respondents with probable mild cognitive impairment. We tried to anticipate the likely cost or burden of these possibilities alongside the likely payoff for research. Table 1 provides a summary of each recommendation. We provide more justification below, roughly in order from easiest to most burdensome or expensive approaches.

Table 1. Enhancements to Strengthen HRS Core Cognitive Content

	Possible Modifications	**Cost/Burden**	**Payoff**
1	Release additional interview metadata predicting cognitive scores	Low	Moderate
2	Incorporate passive follow-up data on those lost to active follow-up (dementia diagnoses from proxy report, Medicare or death certificate) and provide imputed values for missing cognitive scores	Low	Moderate
3	Stagger first cognitive assessments for new enrollees	Low	Moderate
4	Interview proxy informants for some respondents who are also able to complete core assessments themselves	Moderate	Moderate
5	Add digit span and animal naming to core interviews	Moderate	Moderate
6	Enhance literacy/premorbid cognition measures in core interviews	Moderate	High
7	Prioritize HCAP or ADAMS comprehensive assessments for racial/ethnic minorities	Moderate?	High
8	Request additional data linkages for premorbid cognition (e.g., school records)	High	High
9	Add detailed cognitive follow up of all cognitively impaired	High	Moderate
10	Add neuroimaging or neuropathology components for selected subsamples	Very high	Very High
11	Implement web based or passive cognitive assessments, potentially with a measurement burst design	Uncertain, likely high	Moderate

1. Release additional variables that account for within-person variance in cognitive test performance, e.g., an interviewer identifier (anonymized). Measurement error is a major challenge in all cognitive assessments and possibly more severe in HRS than many other cohorts. One approach to addressing measurement error is to identify other (non-cognitive) sources of variance in test performance. A simple example of a possibly strong predictor is the interviewer (it is well known, although perhaps rarely published, that some interviewers naturally elicit stronger performance from the individual taking the cognitive test). Another important predictor might be whether the person was the first or second person in the household to take the cognitive assessment. Time of day may also influence test performance. These types of variables are presumably recorded but are not available in the HRS public use data set. The relevance of these variables for test performance could be evaluated easily by someone with access to the HRS test metadata. If the variables predict cognitive test scores, they could be released to improve researchers' ability to model cognitive change, with no additional cost or participant burden.

2. Improve information on people who no longer participate in cognitive assessments. We now have compelling evidence that selection out of the sample is a major source of bias in cognitive aging studies, and that incorporating other sources of information qualitatively changes the results of at least some studies[11, 25, 26]. Other major cohorts, including the Atherosclerosis Risk in Communities study, have found that imputation of missing values based on supplementary information on diagnosis is possible and valuable[27]. Various analytic strategies, including inverse probability weighting on observed predictors of dropout, will succeed only to the extent that observed covariates account for the links between cognition and dropout. Previous work finds that structural brain changes as measured on MRI predict dropout independently of observed covariates, even observed cognitive scores[28], so it is doubtful that inverse probability weighted methods can fully account for selective dropout. Another source of information on cognition of participants who leave the study would be dementia diagnoses, from Medicare and from death records. Creation of an imputed cognitive score based on these sources would be a major resource for all researchers studying cognitive aging in HRS. Even for individuals who remain in the sample, additional imputations might be helpful.

3. Stagger first exposure to cognitive assessments for new enrollees. This change is suggested for the HCAP but could also be implemented for new enrollees regardless of decisions on the timing of HCAP. Practice effects are easily quantified by using randomized timing for introduction of the first testing occasion. Given the size of HRS cohorts, this can be done without too much cost, and randomizing the first test occasion opens up space in the HRS survey to address other topics for part of the sample, so there is little net scientific loss. Subsamples of new cohorts could be randomly selected for delayed cognitive assessments, allowing for evaluation of practice effects. The size of the subsample should be based on the magnitude of rate of change we wish to detect. If our goal is to place about 0.10 SD window of precision on the estimated practice effect, we need about 1,500 individuals randomized to a delayed cognitive battery. The largest practice effect bump is generally at the first retest, although subsequent retest bumps may contribute a little more.

4. Introduce proxy measures for a sample of individuals who are also directly assessed. HRS already includes proxy cognition measures, but these are only available for individuals who do not participate in direct assessments. This makes it challenging to place the cognitive measures on the same scale as the direct assessments, because there are no people who have both direct assessments and proxy reports. Wu et al placed the proxy responses on the same scale as the direct assessments using the ADAMS sample, and showed the proxy responses were important predictors of function. Unfortunately, the sample was small, the ADAMS battery occurred months or years after the core interview, and estimates were correspondingly imprecise[6]. A simple improvement in the value of the data would be to administer the proxy cognitive assessments to some proxies for a random sample of participants who also complete direct assessments (weighted to overrepresent low performers). This would strengthen the credibility of the imputation model linking the proxy reports to the missing cognitive values. It is likely not necessary to use the full 16-item IQCode, because responses to these items are highly correlated.

5. Add digit span and animal naming to the core assessments. In total, these two assessments add 4 minutes, but have substantial additional variance. If we conceptualize a general cognition construct that influences all cognitive assessments (ignoring for the moment

specific domains of function), given the average correlation between test scores of r=.6, we expect adding these two brief assessments would improve reliability of the overall estimate to 0.86[1].

6. Add a strong literacy measure (e.g., NART) and ideally numeracy measures to the core content. This may add some time but will probably be useful for countless research areas, especially determinants of cognitive aging but also research on decision making, financial planning, health care access, and other topics. The key reason we recommend this is that a literacy measure will help substantially account for pre-morbid variance in cognitive test performance, which should dramatically improve our ability to detect within-person change in cognition.

7. Prioritize more detailed assessments on racial/ethnic minorities to augment the ADAMS data. There is a dearth of good cognitive assessments on older African Americans, Latinos, and Asian Americans. Nearly everything we know is from a handful of cohorts, and these cohorts are all geographically localized. For example, two major sources of data on cognitive aging in Latinos are the WHICAP and SALSA cohorts. Latinos in WHICAP are primarily from the Caribbean, whereas Latinos in SALSA are primarily Mexican American (and there is no non-Latino comparison for SALSA). It is unclear whether the differences we see between Latinos in WHICAP and SALSA reflect differences in Mexican versus Caribbean origin Latinos, differences in the fielding of the surveys, or other factors. Similar challenges prevail when trying to understand black-white disparities in cognitive outcomes. HRS, by virtue of its national sample, could substantially help address this type of evidence gap if comprehensive cognitive assessments were prioritized for racial/ethnic minorities. An important challenge in understanding cognitive aging in diverse populations is selecting measures that are not sensitive to literacy, cultural norms, or test familiarity. Assessments which are robust to such differences are of course of particular value, but current evidence suggests standard neuropsychological assessments perform differently across such dimensions. Exploring the performance of novel measures, e.g., tablet based measures that are less language dependent, could be a useful module fielded for a subgroup of HRS participants.

8. Ask participants if they would allow data linkage to other records (e.g., school records). Such linkages would be valuable even if only feasible for a portion of the sample. Administrative data linkages have multiplied the value of HRS data. There are many sources of administrative data that could speak to premorbid cognition, most obviously school records. If participants consent to such data linkages, these could be used to understand premorbid resources that influence cognitive aging and help account for variance in cognitive function, without additional participant burden.

9. Frequency of design follow-up—increase after age or based on risk model to better capture changes in health and functioning. As with the adaptive HCAP protocol, the core battery could also be implemented with an adaptive design, providing more frequent assessments for impaired or rapidly deteriorating individuals.

10. Neuroimaging or neuropathology measures. These would be very high impact additions to HRS, and of particular value because the cohort is so well-characterized, diverse, and representative. Because of prior investments, each dollar invested in adding neuroimaging on

[1] Spearman-Brown prophecy formula

HRS participants may return more in terms of research impact than investments in neuroimaging new samples. Cost and participant burden may preclude such additions on the whole sample, but a wisely selected subsample might be feasible and extremely informative. Without making recommendations about specific neuroimaging modalities, desirable features here may differ from typical neuroimaging priorities (which emphasize new, expensive approaches that promise greater sensitivity to future dementia risk). Rather, for HRS, the neuroimaging measures of most value would be prioritized based on: (1) strongest associations with neuropsychological test performance changes; (2) relative simplicity of protocol, so consistent information could be collected at multiple sites across the US; (3) mirroring an imaging protocol in existing multi-site cohorts (e.g., ADNI, or possibly CHS) to enable comparison and alignment with those studies. Studies with huge investments, such as ADNI, also had unknown and likely extremely non-representative selection processes. This has posed a challenge for population health researchers interested in using ADNI data: we have very little information on how ADNI patterns may correspond or differ from what would be seen in a population sample. Parallel neuroimaging data – on even the most basic neuroimaging modality – would be invaluable to help us understand and interpret the data from ADNI. This is a case when a modest investment in HRS could pay off in making other data sources much more informative. Very similar issues relate to neuropathology, although there are fewer cohorts with existing neuropathology data. In particular, it is a national priority to obtain more neuropathology on racial/ethnic minorities. HRS *might* be in a strong position to help achieve this goal, given the long-standing participation. That said, brain donation is a sensitive topic (understandably) and this may not be a feasible goal.

11. Incorporate web-based or computer-interface cognitive assessments, in particular passive measures. An internet-based test battery would permit tailored study protocols and temporal assessments to reliably capture within-person processes and change while minimizing an individual's time involvement in the assessment procedures. Web-based testing may also minimize problems associated with an individual's relocation and help to circumvent the interval censoring problem endemic in typical longitudinal designs. Such a format would allow incorporation of a measurement burst design. Measurement burst designs, by helping participants achieve maximum performance after a burst of closely spaced repeated measures, ameliorate retest effects. This type of design can be seen as a special case of the interrupted time-series design[23], with the distinction that the interruptions are planned and based on theoretical and empirical decisions regarding the within-person variation and change in the outcomes of interest.

Conclusion

HRS can provide a powerful platform for research on cognitive aging, dementia, and AD. The major limitations of the HRS assessments relate to the brevity and measurement error in the assessment protocol, which introduce substantial noise to assessments. Improving dementia diagnosis is not a sufficient solution to this problem, because of the timing and developmental nature of ADAD. Rather, we need more precise longitudinal measures of cognitive change, measured on a continuous scale. Adding the HCAP will substantially enhance the value of the cohort for ADAD research, especially when fielded for racial/ethnic minority participants. Several

design variations can help reduce cost and participant burden while optimizing the value of information from the new assessments. These generally entail increased logistical complexity, and so may need to be accompanied by some extra assistance for data-users. Additional changes to the core cognitive battery, on a spectrum from almost free to quite expensive, could also be valuable, both directly for HRS data users and indirectly by enhancing the value of other major data sources on cognitive aging and dementia.

1. Herzog AR, Wallace RB. Measures of cognitive functioning in the AHEAD Study. Journals of Gerontology Series B, Psychological Sciences & Social Sciences 1997;52:37-48.
2. Lachman M, Spiro A. Critique of cognitive measures in the health retirement study (HRS) and the asset and health dynamics among the oldest old (AHEAD) study. US: National Institute on Aging 2002.
3. Banks J, Oldfield Z. Understanding pensions: Cognitive function, numerical ability and retirement saving. Fiscal studies 2007:143-170.
4. Smith JP, McArdle JJ, Willis R. Financial decision making and cognition in a family Context*. The Economic Journal 2010;120:F363-F380.
5. Weir DR, Wallace RB, Langa KM, et al. Reducing case ascertainment costs in US population studies of Alzheimer's disease, dementia, and cognitive impairment-Part 1. Alzheimer's & Dementia: The Journal of the Alzheimer's Association 2011;7:94.
6. Wu Q, Tchetgen EJT, Osypuk TL, White K, Mujahid M, Glymour MM. Combining direct and proxy assessments to reduce attrition bias in a longitudinal study. Alzheimer disease and associated disorders 2013;27:207.
7. Gross AL, Jones RN, Fong TG, Tommet D, Inouye SK. Calibration and validation of an innovative approach for estimating general cognitive performance. Neuroepidemiology 2014;42:144-153.
8. Nguyen T, Tchetgen Tchetgen E, Kawachi I, et al. Instrumental variable approaches to identifying the causal effect of educational attainment on dementia risk. Annals of Epidemiology 2015.
9. Liu SY, Glymour MM, Zahodne LB, Weiss C, Manly JJ. Role of Place in Explaining Racial Heterogeneity in Cognitive Outcomes among Older Adults. Journal of the International Neuropsychological Society 2015;21:677-687.
10. Wang Q, Mejía-Guevara I, Rist P, Walter S, Capistrant B, Glymour M. Changes in Memory before and after Stroke Differ by Age and Sex, but Not by Race. Cerebrovascular Diseases 2014;37:235-243.
11. Weuve J, Proust-Lima C, Power MC, et al. Guidelines for reporting methodological challenges and evaluating potential bias in dementia research. Alzheimer's & Dementia 2015;11:1098-1109.
12. Rast P, Hofer SM. Longitudinal design considerations to optimize power to detect variances and covariances among rates of change: Simulation results based on actual longitudinal studies. Psychological methods 2014;19:133.

13. Liu S, Jones RN, Glymour MM. Implications of lifecourse epidemiology for research on determinants of adult disease. Public health reviews 2010;32:489.
14. Bennett DA, Wilson RS, Schneider JA, et al. Education modifies the relation of AD pathology to level of cognitive function in older persons. Neurology 2003;60:1909.
15. Dufouil C, Alperovitch A, Tzourio C. Influence of education on the relationship between white matter lesions and cognition. Neurology 2003;60:831-836.
16. Jack CR, Knopman DS, Jagust WJ, et al. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. The Lancet Neurology 2013;12:207-216.
17. Braak H, Thal DR, Ghebremedhin E, Del Tredici K. Stages of the pathologic process in Alzheimer disease: age categories from 1 to 100 years. Journal of Neuropathology & Experimental Neurology 2011;70:960-969.
18. Amieva H, Le Goff M, Millet X, et al. Prodromal Alzheimer's disease: successive emergence of the clinical symptoms. Annals of neurology 2008;64:492-498.
19. Laske C, Bateman R, Xiong C, Benzinger T. Clinical and biomarker changes in Alzheimer's disease. The New England journal of medicine 2012;367:2050; author reply 2051-2052.
20. Wilson RS, Li Y, Bienias JL, Bennett DA. Cognitive decline in old age: Separating retest effects from the effects of growing older. Psychology and Aging 2006;21:774-789.
21. Nesselroade JR, McCollam KMS. Putting the process in developmental processes. International Journal of Behavioral Development 2000;24:295-300.
22. Sliwinski MJ. Measurement-Burst Designs for Social Health Research. Social and Personality Psychology Compass 2008;2:245-261.
23. Walls T, Barta W, Stawski R, Collyer C, Hofer S. Timescale-dependent longitudinal designs. Handbook of developmental research methods 2011:46-64.
24. Nesselroade JR. Interindividual differences in intraindividual change. 1991.
25. Mayeda E, Tchetgen Tchetgen E, Power MC, et al. A simulation platform to quantify survival bias: an application to research on determinants of cognitive decline. American Journal of Epidemiology 2016;Forthcoming.
26. Leffondré K, Touraine C, Helmer C, Joly P. Interval-censored time-to-event and competing risk with death: is the illness-death model more accurate than the Cox model? International journal of epidemiology 2013;42:1177-1186.
27. Gottesman RF, Schneider AL, Albert M, et al. Midlife hypertension and 20-year cognitive change: the atherosclerosis risk in communities neurocognitive study. JAMA neurology 2014;71:1218-1227.
28. Glymour MM, Chêne G, Tzourio C, Dufouil C. Brain MRI markers and dropout in a longitudinal study of cognitive aging The Three-City Dijon Study. Neurology 2012;79:1340-1348.