Research Article

The added value of metadata on test completion time for the quantification of cognitive functioning in survey research

Emma Nichols[1,2] ⓘ, Michael Markot[1], Alden L. Gross[3] ⓘ, Richard N. Jones[4] ⓘ, Erik Meijer[1] ⓘ, Stefan Schneider[1,2,5] and Jinkook Lee[1,6]

[1]Center for Economic and Social Research, University of Southern California, Los Angeles, CA, USA, [2]Leonard Davis School of Gerontology, University of Southern California, Los Angeles, CA, USA, [3]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA, [4]Department of Psychiatry and Human Behavior, Warren Alpert Medical School, Brown University, Providence, RI, USA, [5]Department of Psychiatry, University of Southern California, Los Angeles, CA, USA and [6]Department of Economics, University of Southern California, Los Angeles, CA, USA

Abstract

Objective: Information on the time spent completing cognitive testing is often collected, but such data are not typically considered when quantifying cognition in large-scale community-based surveys. We sought to evaluate the added value of timing data over and above traditional cognitive scores for the measurement of cognition in older adults. **Method:** We used data from the Longitudinal Aging Study in India-Diagnostic Assessment of Dementia (LASI-DAD) study ($N = 4,091$), to assess the added value of timing data over and above traditional cognitive scores, using item-specific regression models for 36 cognitive test items. Models were adjusted for age, gender, interviewer, and item score. **Results:** Compared to Quintile 3 (median time), taking longer to complete specific items was associated ($p < 0.05$) with lower cognitive performance for 67% (Quintile 5) and 28% (Quintile 4) of items. Responding quickly (Quintile 1) was associated with higher cognitive performance for 25% of simpler items (e.g., orientation for year), but with lower cognitive functioning for 63% of items requiring higher-order processing (e.g., digit span test). Results were consistent in a range of different analyses adjusting for factors including education, hearing impairment, and language of administration and in models using splines rather than quintiles. **Conclusions:** Response times from cognitive testing may contain important information on cognition not captured in traditional scoring. Incorporation of this information has the potential to improve existing estimates of cognitive functioning.

Keywords: Cognition; timing; paradata; measurement; precision; dementia

(Received 19 July 2024; final revision 13 December 2024; accepted 16 December 2024)

Introduction

In large-scale surveys with objective measures of cognitive functioning, constraints on time and resources underscore the importance of optimizing the quantification of cognition given available information. Increased adoption and use of computer-assisted personal interview (CAPI) methods for cognitive testing has led to increases in the availability of new types of survey metadata, including information on the time spent on cognitive testing (Banerjee et al., 2020; Humphreys et al., 2017; Shega et al., 2014). The use of speeded tests, such as the Trail-Making Test, to assess cognition supports the general notion that response times on cognitive tests contain meaningful information about cognitive functioning (Bowie & Harvey, 2006; Salthouse, 1996). However, in large-scale community-based survey research on cognitive aging and dementia, this information is not typically used beyond general data monitoring procedures and quality control.

A substantial body of evidence from psychology and educational testing provides further support for the utility of information on timing and suggests a tradeoff between timing and accuracy or score; those with the fastest times may be guessing or otherwise sacrificing accuracy for speed (Kyllonen & Zu, 2016; Wickelgren, 1977). Longer response times may indicate more careful attention to the question or may be a sign of increased struggle in understanding and answering a given question. The complexities of the combined effects of guessing, careful consideration of questions, and observed difficulties with challenging questions likely help explain observed nonmonotonic, nonlinear relationships between response times and ability (Chen, De Boeck, Grady, Yang, & Waldschmidt, 2018; Dodonov & Dodonova, 2012). Evidence also suggests that the impacts of the different factors affecting response times (e.g., guessing, processes around critical reasoning) may vary by item type (DiTrapani et al., 2016; Goldhammer et al., 2014; Hess et al., 2013). For example, one study of items on the Dental Admission Test found that the maximum point on the curve describing the relationship between score and response time (i.e., the time taken associated with the highest

ability level) was earlier for knowledge-based items than items requiring controlled processes (Chen et al., 2018).

Despite robust existing research on response times in psychology and educational testing, relatively few studies have examined the use of information on response times from non-speeded tests to inform the estimation of cognitive functioning and classification of mild cognitive impairment (MCI) and dementia in older adults in large-scale population-based research. Issues associated with challenges with measuring cognitive functioning in participants' homes in large-scale survey research also raises questions about the potential value of timing data in this context. However, there is growing evidence that response times on non-cognitive survey questions are associated with future cognitive functioning, MCI, and dementia (Junghaenel et al., 2023; Schneider et al., 2023; Seelye et al., 2018), which may be useful for the estimation of cognitive functioning in surveys without formal cognitive assessments. However, in studies with objective cognitive testing, timing data may also be used to enhance the available information on cognition from test scores.

To understand the potential benefit of using timing data, it is first necessary to assess whether information on timing from cognitive testing provides useful information independently of the already available information on performance from cognitive test scores. One prior study found that longer and more variable response times for items included in the Montreal Cognitive Assessment (MoCA) were significantly associated with concurrent and future cognitive functioning (Sanders et al., 2024); however, the brevity of the MoCA instrument precludes examination of item-level heterogeneity, and nonlinearities in associations were not considered. Additionally, despite increasing research interest in cross-national and global research on cognitive aging and dementia, this work and other prior research on timing has largely been conducted in the United States. Differences in testing environments at respondent's houses, differences in the distribution and range of socioeconomic status across respondents, and cultural perceptions of time in low- and middle-income countries (LMICs) may influence observed findings.

We sought to build on this existing evidence by using data from the Longitudinal Aging Study in India Diagnostic Assessment of Dementia (LASI-DAD) study to examine the added value of timing data from cognitive testing for the assessment of cognitive functioning in large-scale survey research in the Indian context. We leveraged the large cognitive battery included in the LASI-DAD study to examine associations between cognitive functioning and time spent on cognitive test items across 36 items, considering item-level heterogeneity and allowing for nonlinearities in associations. We also evaluated associations with MCI and dementia as key clinical outcomes.

Method

Sample

LASI-DAD is a nationally representative sample of 4,096 adults 60 years and older in India (Lee, Banerjee, Khobragade, Angrisani, & Dey, 2019; Lee, Khobragade, et al., 2020). LASI-DAD was designed as a sub-study of the broader Longitudinal Aging Study in India (LASI) study with focus on the assessment of cognition and dementia. LASI-DAD uses stratified sampled procedures to select LASI participants from 18 states and union territories with equal numbers of those at high and low risk of cognitive impairment based on cognitive testing in LASI. We used data from $N = 4,091$ respondents with available information on the timing of cognitive

tests from the CAPI survey instrument. All participants gave informed consent for participation (either written or thumbprint). We obtained ethics approval from the Indian Council of Medical Research (2202-16741/F1) and all collaborating institutions and research was conducted in accordance with the Helsinki Declaration.

Cognitive functioning

The LASI-DAD study administered the Harmonized Cognitive Assessment Protocol (HCAP) battery of cognitive tests and informant reports, which has been described in detail elsewhere (Langa et al., 2020; Lee, Khobragade, et al., 2020). The cognitive battery was designed to take approximately 1 hour, and included tests assessing orientation, memory, language, executive functioning, and visuospatial functioning (Supplementary Materials 1). Some adaptations to the original HCAP battery were necessary to ensure cultural appropriateness and adequate performance in a population with low education and literacy (Banerjee et al., 2020). The battery was translated into 12 languages and was administered in each participant's preferred spoken language. A continuous measure of cognitive functioning based on all available cognitive tests was previously estimated using confirmatory factor analysis, and was shown to perform equivalently across language of administration (Gross et al., 2020). The measure is scaled to have a mean of 0 and standard deviation of 1 in the full LASI-DAD sample.

Cognitive test timing

The time spent on each cognitive task was captured by the CAPI survey system as the time spent on each screen. Each screen typically included a combination of related instructions and one or more items. For cognitive tests that assess a specific cognitive process, and which are not typically disaggregated at the item level (e.g., Raven's progressive matrices), time spent on multiple individual items belonging to the same cognitive test was aggregated to the test level. To simplify the text in the present study, we use the term "item" to refer to the mixture of both items and tests considered (test-level scores and timing from tests of specific cognitive processes and item-level scores and timing from more heterogeneous test batteries such as the Hindi Mental State Examination) (Ganguli et al., 1995).

We excluded the first items of the judgement and problem-solving item set, Token test, and Raven's progressive matrices, as these items appeared on the same screen as large amounts of introductory text and instructions and it was therefore impossible to disentangle the time spent on instructions from the time spent answering these items. Items for orientation to time (what is the month, year, day of week, and date) were presented on the same screen and were therefore summed together. We excluded items conducted with pencil and paper (write sentence, copy pentagons, constructional praxis, clock drawing), as timing data was not collected. We also excluded delayed story recall, as each story point was assessed on a separate screen; due to the brief nature of the items in contrast to the time taken to switch screens, the captured timing measure would likely be less reflective of respondent ability. Finally, we excluded the item for floor of building as it was not administered to over 75% of respondents. It is important to acknowledge that time as captured by the CAPI instrument includes the interviewer's delivery of the question and could also include any interruptions to the interview that may have occurred while on a specific screen.

To reduce the number of outliers due to testing interruptions, we examined the distributions of timing data and set extreme observations to missing. The items varied in the skewness of the observed timing data with a bimodal distribution of variable skewness across tests. Therefore, we defined extreme as the 99th percentile of the data distribution for items with low skew, and the 98th percentile of the data distribution for items with high skew (Supplementary Materials 2). Because this method was designed to be conservative in removing data, we selected analytic approaches (analysis of quintiles, splines with boundary knots) that are robust to outliers at the extremes.

We divided the timing data for each item into quintiles for use in primary analyses. In addition to considering timing data on individual test items, we also calculated quintiles of an overall timing variable to characterize broad patterns in the timing data, considering the time spent on all items. The overall timing variable was calculated as the sum of time spent across all items considered in the analyses.

Cognitive impairment status

We used Clinical Dementia Rating (CDR) scale classifications from an online clinical consensus process. We compared CDR = 0 (no dementia), CDR = 0.5 (questionable dementia or MCI, here referred to as MCI), and CDR ≥ 1 (dementia). Details on the process for the online adjudication procedure are available elsewhere (Lee, Ganguli, et al., 2020). Because CDR adjudication was not completed for participants in Phase 1 of data collection, analyses using CDR classifications were conducted on a subset of respondents (N = 2,525).

Other measured covariates

We also used self-reported information on age, sex/gender (hereafter gender), education (none/less than secondary/secondary or higher), marital status (married or partnered/other), and hearing impairment (whether the respondent ever had hearing/ear-related problems or conditions). Rural/urban residence was determined based on the classification of locations in the 2011 Census.

Statistical analysis

We examined the characteristics of the sample by demographic variables, both overall and by quintiles of the overall timing variable (timing summed across items). We also assessed the means and distributions of cognitive scores across quintiles of the overall timing variable.

Primary analyses consisted of regression models for each item (or set of items appearing together on the CAPI screen) to estimate associations between general cognitive functioning, the dependent variable, and indicators for quintile of time spent on the test item, the independent variable. All models adjusted for age, gender, interviewer, and score on the specific item. By adjusting for score on the specific item, models capture the added value of timing data for the measurement of cognitive functioning, beyond traditional test scores. We used five quintiles to ensure that models had sufficient flexibility to capture nonlinearities while prioritizing the interpretability of estimates. To ensure findings were not sensitive to our choice of the number of bins included, we also estimated item-specific models using flexible splines for item timing to assess the consistency of findings. We used restricted cubic splines with 3 degrees of freedom and boundary knots at the 5th and 90th percentiles, which constrain the tail segments to be linear to prevent potential outliers from having undue influence. We standardized all continuous item-specific timing data prior to estimating spline models to facilitate comparisons across items. To evaluate the statistical significance of comparisons between different points on the estimated spline terms (analogous to the comparison of quintiles in primary analyses), we used a non-parametric bootstrap (details in Supplementary Materials 3).

We then assessed whether item difficulty could explain heterogeneity in findings across items by numerically and visually evaluating correlations between item difficulty and the item-level differences in cognitive functioning between the third and first quintiles of time spent on a specific item. We used item response theory methods to quantify the average item difficulty across item-specific thresholds (additional details in Supplementary Materials 4). Item difficulties for each item are in Supplementary Materials 5.

To evaluate the utility of information on timing for classification purposes, we first estimated multinomial logistic models of CDR MCI and dementia on quintile of time spent on each item. Models were analogous to those used in the primary analyses; however, we used multinomial logistic regression given the categorical nature of the CDR outcome. To further assess whether data on timing may contain information on incipient impairments that are present before changes can be observed in cognitive test scores, we estimated the discordance between cognitive functioning and timing data. Specifically, for each item we calculated the following proportion:

$$\frac{N \text{ with cognition above the mean and timing data suggestive of cognition below the mean}}{N \text{ with cognition above the mean}}$$

We defined timing data suggestive of cognition below the mean as timing data for which the observed time spent was more extreme than the time for which the upper bound of the mean predicted cognitive functioning crossed 0. We only considered times in the tails of the distributions, as the timing data in these regions showed the largest associations with cognitive functioning in primary analyses. These criteria were designed to identify respondents who scored well on cognitive testing, but who, based on timing data, completed cognitive testing in a manner that would suggest lower cognitive functioning. For these respondents, timing data may indicate subtle cognitive deficits that are unobservable using the existing cognitive tests.

We conducted sensitivity analyses to evaluate the stability of our primary results to additional adjustment for language of administration, education, urbanicity, and hearing impairment. We considered adjustment for language of administration as a sensitivity analysis because language was somewhat collinear with interviewer and added a significant number of terms to models, reducing precision. Adjustments for education, urbanicity as a proxy for noise and interruptions, and hearing impairment were considered sensitivity analyses because while these factors may impact speed independently of cognitive functioning through mechanisms such as familiarity with test taking or ability to hear instructions better, evidence suggests these factors may also causally affect cognitive functioning and variation in cognition due to these potentially causal links should not be adjusted away. We also estimated models with executive functioning as the cognitive outcome, as attention/speed was considered a component of executive functioning in factor analyses (Gross et al., 2020) and therefore timing data may be more closely related with executive functioning than general cognitive functioning.

Table 1. Demographic characteristics and cognitive scores in the diagnostic assessment of Dementia for the longitudinal aging study in India (LASI-DAD) (*N* = 4,091), stratified by quintile of overall time taken on cognitive tests. Proportions and totals are shown for binary and categorical variables; means and IQR's are shown for continuous variables

		Quintile of overall time on cognitive items				
Variable	Overall	Quintile 1 (fastest)	Quintile 2	Quintile 3 (median)	Quintile 4	Quintile 5 (slowest)
Number of participants	4091	819	819	817	818	818
Age, years, mean (IQR)	69.7 (64–74)	69.7 (63–74)	69.0 (64–73)	69.7 (64–74)	69.8 (64–74)	70.4 (65.0–75.0)
Female gender, % (N)	53.9 (2204)	54.8 (449)	54.2 (444)	48.3 (395)	55.3 (452)	56.7 (464)
Rural, % (N)	61.9 (2533)	58.4 (478)	61.9 (507)	64.5 (527)	61.6 (504)	63.2 (517)
Married/ Partnered, % (N)	65.2 (2669)	63.4 (519)	67.9 (556)	67.6 (552)	64.2 (525)	63.2 (517)
Education, % (N)						
None	49.0 (2006)	50.9 (417)	39.3 (322)	43.5 (355)	52.2 (427)	59.3 (485)
Less than secondary education	34.0 (1389)	30.4 (249)	36.5 (299)	37.6 (307)	32.9 (269)	32.4 (265)
Secondary or higher	17.0 (696)	18.7 (153)	24.2 (198)	19.0 (155)	14.9 (122)	8.3 (68)
General cognition score, mean (IQR)	−0.01 (−0.70–0.67)	−0.09 (−0.83–0.66)	0.20 (−0.57–1.02)	0.13 (−0.55–0.84)	−0.05 (−0.73–0.51)	−0.25 (−0.80–0.23)
Executive functioning score, mean (IQR)	−0.01 (−0.71–0.66)	−0.06 (−0.80–0.65)	0.18 (−0.58–0.94)	0.12 (−0.54–0.80)	−0.07 (−0.73–0.54)	−0.24 (−0.79–0.26)

All analyses were conducted in R version 4.2.2.

Results

Sample characteristics

The LASI-DAD sample used for this analysis included 4,091 respondents with a mean age of 69.7 years (Interquartile Range [IQR] 64–74); 53.9% were women (Table 1). The majority (61.9%) of respondents lived in rural settings, and almost half (49.0%) had no formal education. There were some differences in demographics by quintile of overall time spent on items; for example, respondents in Quintile 2 were on average younger (69.0 years) than those in the fastest (Quintiles 1; 69.7 years) or slowest (Quintile 5; 70.4 years)) quintiles. A smaller proportion of respondents in Quintile 2 had no education (39.3%) compared to Quintiles 1 (50.9%) and 5 (59.3%).

Univariate associations between cognition and overall time on cognitive items

The distributions of cognitive functioning were largely overlapping, comparing across quintiles of overall time spent on cognitive tasks, but there was a clear pattern when examining means of distributions (Table 1, Supplementary Materials 6). The overall pattern was nonmonotonic: increasing between Quintiles 1 (Mean: −0.09 SD) and 2 (0.20 SD), and then decreasing across the remaining 3 Quintiles (3: 0.13 SD; 4: −0.05 SD; 5: −0.25 SD). A similar pattern was observed for executive functioning.

Item-specific associations between timing and cognitive functioning

To examine item-specific associations, we assessed models in which we predicted general cognitive functioning (the factor score) as a function of an item score and the time spent on the particular item, again categorized into quintiles, with Quintile 3 the reference category. Compared to Quintile 3, respondents who spent slightly less time on cognitive test items (Quintile 2) tended to have higher cognitive functioning on average, holding their observed scores on the specific item constant (Figure 1). Of the 36 items considered, 13 (36%) had significantly positive associations, suggesting that

slightly faster times to complete these 18 items is associated with better cognition; only 1 had a significantly negative association.

In contrast, respondents who spent more time on items compared to those in Quintile 3 tended to have lower cognitive functioning holding observed scores constant. Results were stronger for Quintile 5 (24 items with negative associations, 0 positive) than Quintile 4 (10 items with negative associations, 1 positive). For Quintile 1, results were less consistent, with time spent on 7 items having positive and 11 negative associations with cognitive functioning, holding score constant.

Overall findings were consistent in models with executive functioning as the outcome instead of general cognitive performance, and in models controlling for either education, urbanicity, language of administration, or hearing impairment (Supplementary Materials 7–11). Use of flexible cubic splines in lieu of quintiles also yielded results with consistent overall findings: for most items, predicted cognitive functioning given timing data suggested low levels of cognitive functioning for the shortest response times and as the time spent on items increased, predicted cognitive functioning then peaked quickly and tapered off after 0.5 standard deviation units of response time with a long right tail (Figure 2). Comparisons between predictions at different points on the curves were largely statistically significant (Supplementary Materials 3). We observed the most heterogeneity across items in the left tails. This aligns with our findings regarding Quintile 1 in the main analyses: for some items, after adjusting for item score, answering very quickly was associated with high cognitive functioning, while for other items this is associated with low cognitive functioning after adjusting for item score.

Item difficulty and item-level differences

Item difficulty was negatively correlated with coefficients for Quintile 1 ($p = 0.002$), explaining some of the observed heterogeneity (Supplementary Materials 12). A negative correlation suggests that for easier items with lower difficulty, coefficients for Quintile 1 are more likely to be positive, suggesting that quicker response times are associated with higher cognition, controlling for item score. However, for more difficult, complex items (higher difficulty), coefficients for Quintile 1 are more likely to be negative, suggesting that quicker response times are associated with lower
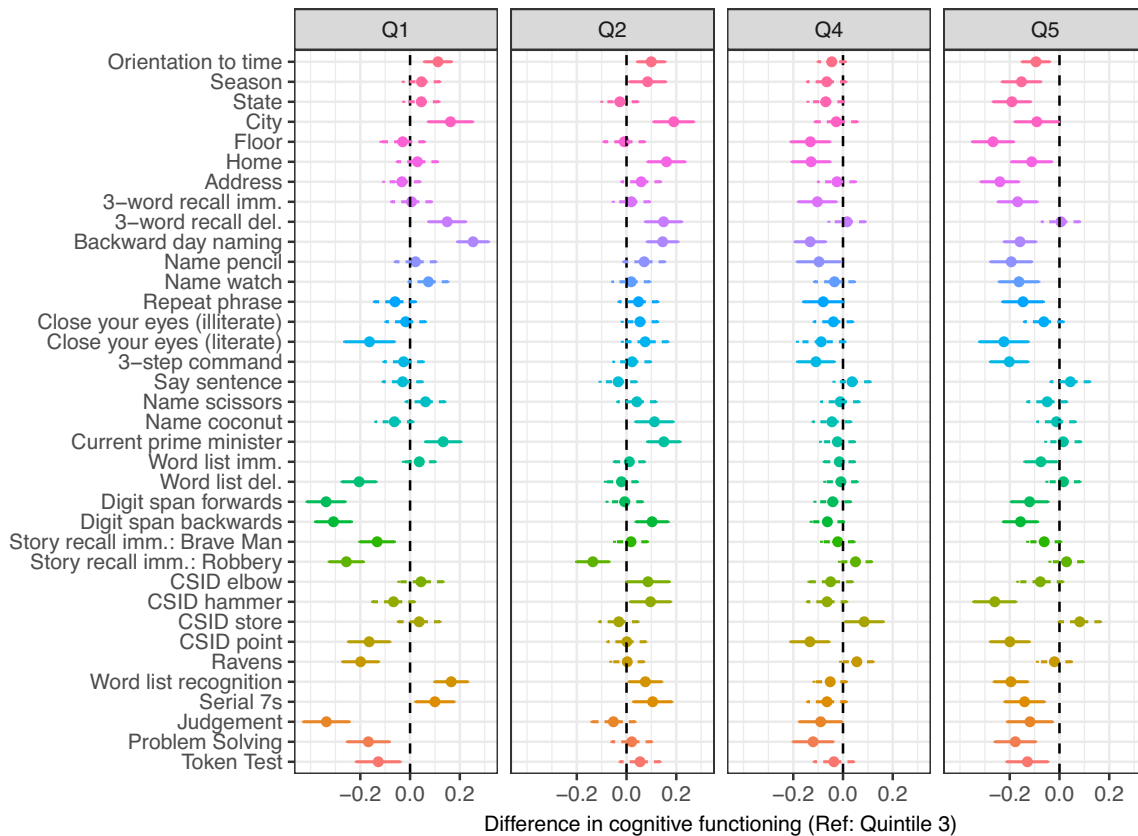
Figure 1. Differences in mean cognitive functioning for each quintile of time taken to complete individual cognitive tests compared to Quintile 3; positive coefficients suggest membership in the quintile is associated with better general cognitive functioning than Quintile 3 on average. Estimates were derived from item-specific linear regression models of general cognitive functioning regressed on quintiles of time taken on each specific subtest, controlling for age, gender, interviewer, and score of the subtest. Uncertainty intervals show 95% confidence intervals; lines are solid if the 95% confidence interval does not cross 0 and dotted if it does. colors are used to help differentiate the estimates.

cognition, controlling for item score, potentially due to the respondent guessing or giving up.

Item-specific associations between timing and MCI and dementia

Associations between item-level timing and odds of MCI or dementia were consistent with findings using continuous cognition, albeit somewhat weaker (Figure 3). Of the 36 items, slower test performance was associated with elevated odds of MCI for 6 and 14 items for Quintiles 4 and 5, respectively. Similarly, slower test performance was associated with elevated odds of dementia for 4 and 13 items for Quintiles 4 and 5, respectively. Although observed effect sizes were larger for dementia, the number of statistically significant associations in the expected direction (based on primary analyses) was similar between MCI and dementia for Quintiles 4 and 5. However, findings for MCI and dementia diverged for Quintiles 1 and 2, with, for example, 6 significant negative associations and 0 positive associations for MCI and 7 significant positive associations and 0 negative associations for dementia in Quintile 1. This suggests that quick answers were associated with lower odds of MCI but higher odds of dementia, perhaps indicating that the effect of cognitive status on giving up and therefore answering quickly is only present at later stages of disease.

Proportion of respondents with mismatched cognition and timing data

The proportion of respondents who had above average cognitive functioning (factor scores above 0) but timing data suggestive of below average cognitive functioning varied between items from 43% (token test) to 0% (serial 7s, 4 other items) (Figure 4, Supplementary Materials 13). The proportion of respondents with discordant cognitive performance and timing data was above 20% for 25/36 items, suggesting a substantial number of individuals with intact cognition as measured by cognitive test scores, but for whom timing data may suggest subtle cognitive deficits.

Discussion

Using representative data on older adults from the LASI-DAD study, we found that data on timing from cognitive tests may be useful in improving or augmenting existing measures of cognitive functioning from large-scale community-based surveys. Analyses highlighted evidence that time spent on items was associated with cognitive functioning and cognitive outcomes after controlling for the traditional score on the same cognitive test. Longer time (than the central quintile) spent on items was associated with lower cognition, whereas shorter time (than the central quintile) spent on items was associated with either higher or lower cognition
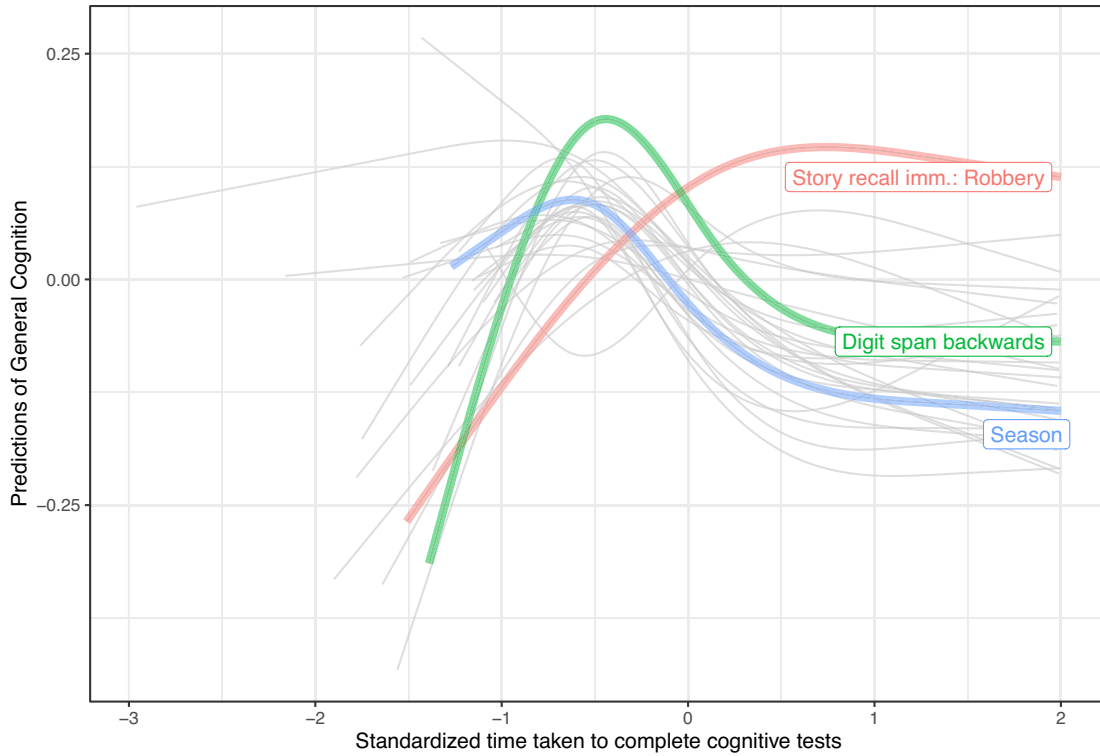
Figure 2. Smooth estimates of predicted general cognitive functioning by item-specific standardized time taken to complete cognitive tests. Estimates were derived from item-specific regression models for the association between general cognitive functioning and time taken on each specific test controlling for age, gender, interviewer, and score of the test. Time spent on each test was modeled using a cubic spline with 3 degrees of freedom and boundary knots at the 5 and 90% percentiles.
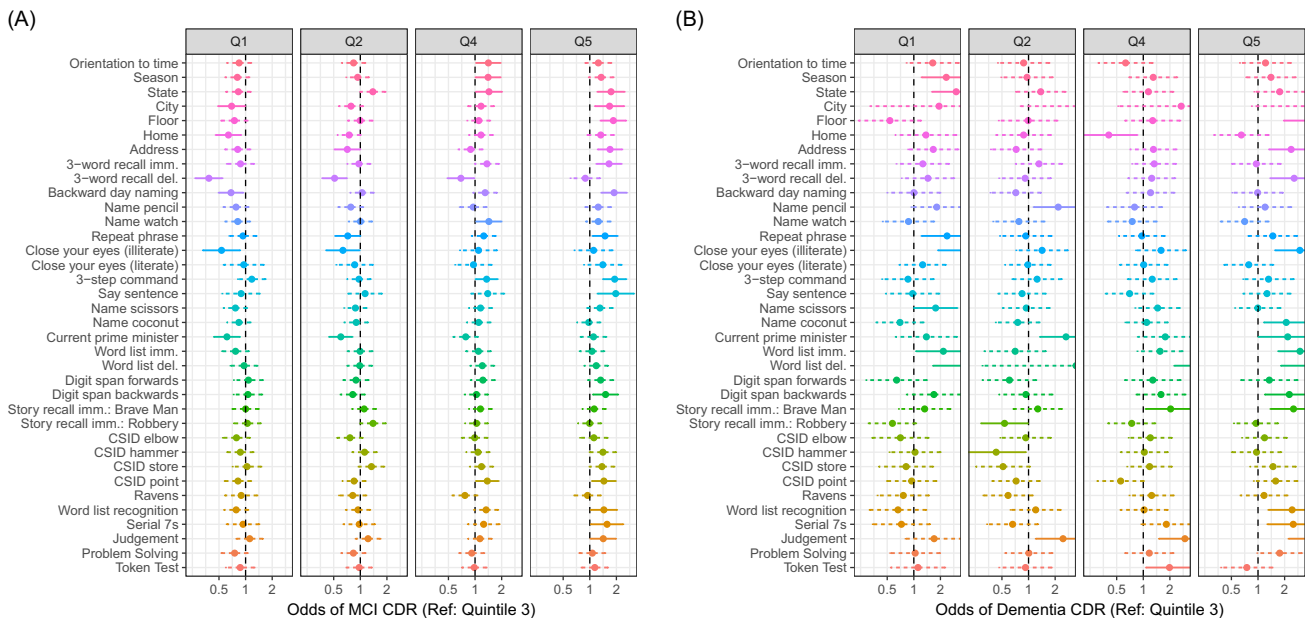

Figure 3. Odds ratios for dementia or mild cognitive impairment for each quintile of time taken to complete a given cognitive tests compared to Quintile 3. Estimates were derived from item-specific multinomial logistic regression models for the association between general cognitive functioning and quintile of time taken on each specific test controlling for age, gender, interviewer, and score of the test. Uncertainty intervals show 95% confidence intervals; lines are solid if the 95% confidence interval does not cross 0 and dotted if it does. colors are used to help differentiate the estimates.
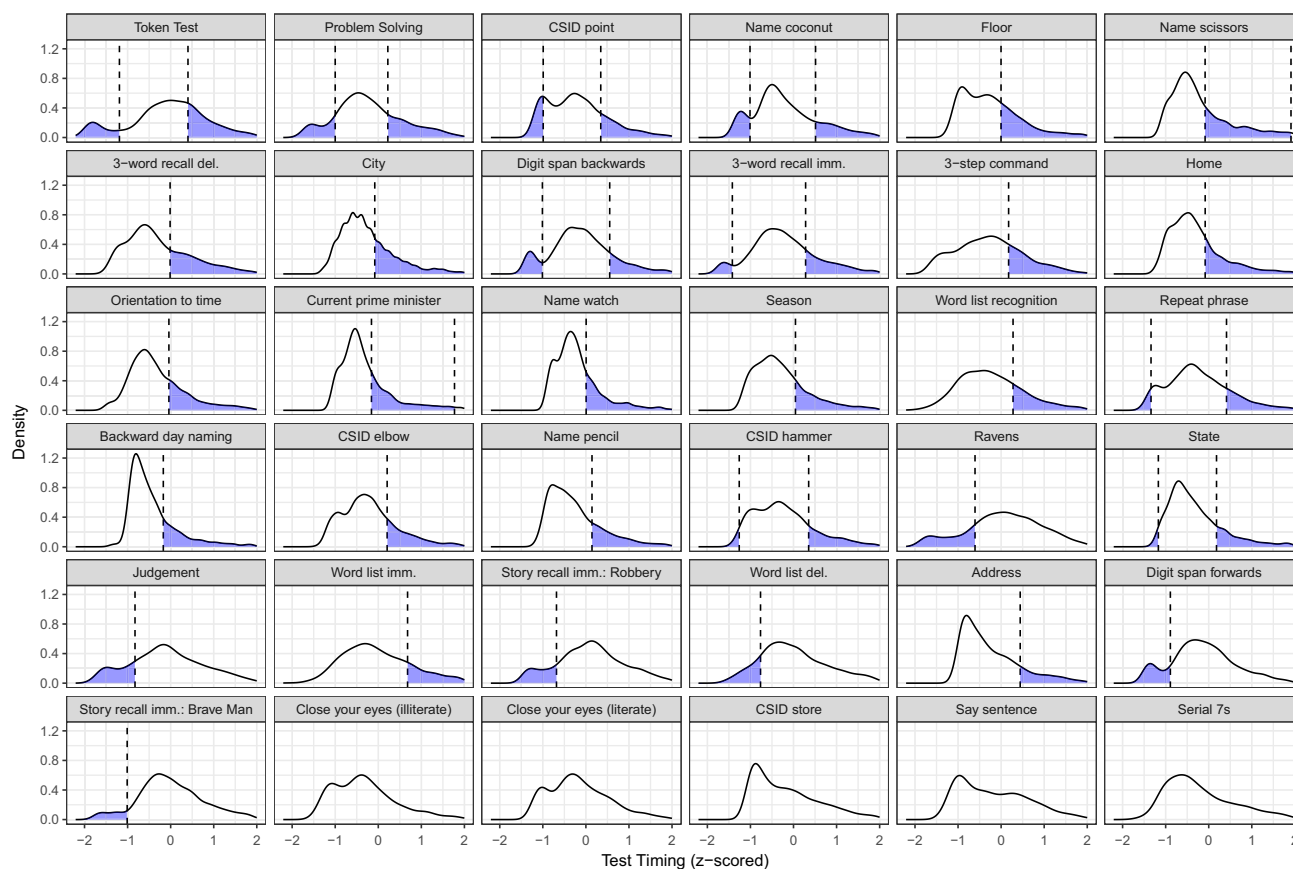
Figure 4. Density plots with shaded areas showing the proportion of individuals with cognitive functioning above mean levels (greater than 0 on the standardized general cognitive functioning score), but timing data suggestive of cognitive functioning statistically significantly below the sample mean. For the five tests with no highlighted regions, there are no standardized response speeds at which someone with above average cognitive functioning would be predicted to have below average cognitive functioning.

depending on item difficulty. For more difficult items, shorter completion times were associated with lower cognition, consistent with guessing or rushing through the interview. In contrast, for easier items, shorter completion times were associated with higher cognition. We also identified respondents with measures of objective cognitive functioning above the mean, but timing data suggestive of cognitive functioning significantly below the mean, which may indicate that timing data could be valuable for the identification of subtle cognitive deficits.

Our findings align with prior results from the National Social, Health and Aging Project (NSHAP), which highlighted initial evidence that information on timing captured using CAPI survey instruments may contain important information on cognition (Sanders et al., 2024). Findings from NSHAP also provided longitudinal evidence that information on timing is associated with future cognitive functioning independently of cognitive test scores. As additional waves of LASI-DAD data become available, these findings should be replicated and extended using the larger included cognitive battery in the LASI-DAD study. Our findings also follow from prior evidence on the potential utility of timing data in related contexts. One prior web-based study of 138 older adults found that total completion time of a novel cognitive assessment tool, including time for instructions, practice items, and test items, was highly predictive of cognitive status, lending initial support for the use of meta-data from cognitive testing in assessing cognition (Dorociak et al., 2021). Our results, in tandem

with results from NSHAP, extend this initial evidence by using larger, population-representative samples with more extensive cognitive batteries administered via a CAPI survey instrument.

In comparison to these prior studies, the length of the LASI-DAD battery allowed for a more detailed examination of item-level heterogeneity. Though we focused on larger patterns of results in overall findings to protect against concerns of type I error given the large number of hypothesis tests and models, identification of items with consistently strong associations across various analyses may highlight specific items for which timing data may be particularly indicative of cognitive status. In current analyses, word list recognition, digit span backwards, and the judgement questions were all significantly associated with multiple cognitive outcomes considered, suggesting timing data from these items may be particularly meaningful.

Other existing literature has focused specifically on response times for non-cognitive survey items (Junghaenel et al., 2023; Schneider et al., 2023; Seelye et al., 2018). These studies have found that the time spent on non-cognitive survey items is predictive of future cognitive functioning, MCI, and dementia (Junghaenel et al., 2023; Schneider et al., 2023; Seelye et al., 2018). Findings from these studies motivate the need for replication efforts in diverse settings such as LASI and LASI-DAD and have important implications for the potential utility of existing data sources without objective cognitive testing, as response times may allow for estimation of cognition in these studies. However, beyond this potential

application, findings also more broadly support the notion that information on the time spent on tasks requiring cognitive effort yields meaningful information about cognition.

Our analyses indicated not only that response times from cognitive testing were associated with cognitive functioning, but that they uniquely contributed to explaining underlying cognitive functioning after accounting for scores from standard scoring procedures. These findings suggest that use of information on time spent on cognitive tests may lead to improvements in the precision and quality of derived measures of cognitive functioning based on survey data. Researchers in the field of educational testing have previously developed combined models that either incorporate response times into item-level item-response theory models (Wang & Hanson, 2005) or into joint models that estimate both ability and speed as correlated latent traits (van der Linden, 2008). The correlated traits model was shown to improve precision in a test of fluid reasoning administered to a highly educated population, such that the battery could be reduced by 33% without a corresponding reduction in precision through the incorporation of information on response times (Bertling & Weeks, 2018). The current results suggest that the use of such joint models in cognitive aging research has the potential to improve measurement and should be explored in future work.

Our findings of discordance between objective testing and implied cognition based on timing data suggest that information on timing may lend insights into subtle cognitive deficits that may not be apparent through objective cognitive testing. Of those with estimates of cognitive functioning above the mean, the proportion of respondents with discordant scores on objective testing and timing data on the three items identified with strong associations between timing data and cognitive outcomes was 28% (digit span backwards), 23% (word list recognition), and 21% (judgement questions), indicating sizable proportions for these tasks with strong associations. The notion that this observed discordance may signal subtle cognitive deficits is supported by findings from a prior study of survey meta-data from non-cognitive tests, which found that individuals with incident MCI had increased completion times one year prior to formal diagnoses compared to those who remained cognitively intact (Seelye et al., 2018). Other research has identified different subtle cognitive changes occurring prior to MCI diagnosis, including impaired response inhibition and differences in the patterns of response to neuropsychological tests (Schmid, Taylor, Foldi, Berres, & Monsch, 2013; Wylie et al., 2007). Protective factors, such as education, that may help respondents score well on cognitive testing in the face of subtle deficits may not protect against deficits that could show up in timing data or other metadata.

Taken together, our results and these prior findings indicate that metadata and other assessments of more subtle cognitive problems may be helpful for capturing mild impairments in those with largely intact cognitive functioning. Neuropsychological batteries designed for use in large-scale surveys often include a number of screening tests, such as the Mini-Mental State Examination, which do not have adequate precision for the detection of mild cognitive deficits (Gross, Jones, Fong, Tommet, & Inouye, 2014; Koski et al., 2011). Given the limitations of these batteries, particularly for the assessment of subtle cognitive changes among those with higher levels of cognitive functioning, augmentation of traditional scores with timing metadata could yield important improvements for the estimation of cognition across the entire spectrum of cognitive functioning. Future research should leverage longitudinal data to further evaluate

whether individuals with mismatched data are more likely to decline in future waves, which would provide additional evidence that such discordance indicates early cognitive deficits. In addition to the potential research utility of timing information, the adoption of tablet-based or computerized testing in clinical settings would also facilitate the integration of metadata on time taken to complete cognitive tasks into clinical screening tools for subtle cognitive deficits in clinical settings. The development of such protocols was outside the scope of the current paper but should be the topic of follow-up research.

It is important to consider the ways in which the study context may impact results. Cultural attitudes towards time and experience with high-stakes testing have the potential to impact how respondents react to cognitive tests and prior research has shown this can impact timing data from speeded cognitive tests (Agranovich et al., 2011; Eizaguirre et al., 2020). Social and cultural psychology has suggested LMICs often have a slower pace of life and less emphasis on timeliness compared to high-income contexts (Levine, 2008). Given these patterns, overall lower education levels, and less exposure to high-stakes testing, respondents in LASI-DAD likely have less of an internal sense of time pressure than respondents in high-income contexts. Although differences in culture are important to consider in thinking about the generalizability of findings across settings and replication in other contexts is needed to confirm the consistency of patterns, the consistency of findings regarding the utility of timing data across both high-income contexts and India, despite vast differences in cultural attitudes toward time and educational experiences suggests that these data may be useful across diverse settings.

Additionally, considerations around the extent to which noise and distractions could impact testing are different across study and country contexts. The presence of robust associations based on timing data collected at respondents' homes throughout India, where overall measurement error is likely higher due to higher levels of noise pollution, higher likelihood of interruptions, and a greater number of logistical challenges as compared to lab-based testing or even large-scale surveys in high-income contexts, lends even more strength to observed associations due to the expected attenuation of effects in the presence of this measurement error. Associations may be even stronger in settings with lower noise and fewer interruptions.

Key study strengths include the use of a large, nationally representative dataset with a comprehensive neuropsychological battery, and the use of many supplementary and sensitivity analyses to evaluate the robustness of findings to differences in model form and adjustments for a variety of factors that could impact the time spent on testing. Limitations should also be considered. First, data collection procedures were not optimized for the collection of timing data; information on the time spent on cognitive tests included time for instructions and time due to interruptions in testing and data collection. Respondents did not know in advance that data on response times would be considered by investigators and may have changed their behavior if presented with this information. Despite this lack of optimization, instruction time may contain relevant information; if interviewers slow down for those with cognitive impairments, this may contribute to some of the observed signal in analyses and may be helpful in characterizing cognition. However, to minimize the extent of these concerns in general, we made decisions on a post-hoc, case-by-case basis to exclude items on the same screens as blocks of instructions or groups of items on a single screen and

used data-driven methods to identify outliers most likely due to interruptions. Some residual measurement error in the data due to inaccuracies in the collection of timing data likely remains; our results are conservative in comparison to what would be expected in the presence of less measurement error. With adjustments to the design of the CAPI instrument and interviewer training, it may be possible to reduce measurement error in the collection of timing data and increase its utility.

Second, in item-level analyses a large number of models and hypothesis tests are necessary, leading to some concern about inflated type 1 error. To alleviate these worries and limit concerns around over-interpreting individual findings, we focused our overall conclusions on broad patterns of results, rather than individual findings. Examination of these patterns across the range of items gives confidence that findings are not due to chance. For example, in primary regression models, 24/36 estimates for Quintile 5 were statistically significant and negative while there were no estimates that were statistically significant and positive; however, in contrast, if results were due solely to type 1 error, we would expect approximately 5% (∼2 models) to be statistically significant with a similar number of both positive and negative estimates. Finally, the LASI-DAD sample is currently cross-sectional; therefore, it was not possible to evaluate associations between timing data and future cognitive impairment or clinical status. Additional research is needed to extend the current analysis and evaluate associations with future cognitive outcomes when future waves become available.

Our results advance the evidence supporting the utility of information on the time spent on cognitive tasks, conditional on scores from traditional scoring procedures, even in studies conducted in real-world environments with the potential for interruptions and substantial measurement error. Findings support the integration of timing data from the administration of cognitive testing into the quantification of cognitive functioning in large-scale survey research on cognitive aging. Information may be particularly valuable for the assessment of subtle cognitive changes among those with higher levels of cognitive functioning. Future research is needed to bolster longitudinal evidence and to implement and test methods for the integration of information on timing into models for the quantification of cognitive functioning.

References

Agranovich, A. V., Panter, A. T., Puente, A. E., & Touradji, P. (2011). The culture of time in neuropsychological assessment: Exploring the effects of culture-specific time attitudes on timed test performance in Russian and American samples. *Journal of the International Neuropsychological Society*, *17*, 692–701.

Banerjee, J., Jain, U., Khobragade, P., Weerman, B., Hu, P., Chien, S., Dey, S., Chatterjee, P., Saxton, J., Keller, B., Crimmins, E., Toga, A., Jain, A., Shanthi,

G. S., Kurup, R., Raman, A., Chakrabarti, S. S., Varghese, M., John, J. P., Joshi, H., . . . C., Rajguru (2020). Methodological considerations in designing and implementing the harmonized diagnostic assessment of dementia for longitudinal aging study in India (LASI–DAD). *Biodemography and Social Biology*, *65*, 189–213.

Bertling, M., & Weeks, J. P. (2018). Using response time data to reduce testing time in cognitive tests. *Psychological Assessment*, *30*, 328–338.

Bowie, C. R., & Harvey, P. D. (2006). Administration and interpretation of the trail making test. *Nature Protocols*, *1*, 2277–2281.

Chen, H., De Boeck, P., Grady, M., Yang, C.-L., & Waldschmidt, D. (2018). Curvilinear dependency of response accuracy on response time in cognitive tests. *Intelligence*, *69*, 16–23.

DiTrapani, J., Jeon, M., De Boeck, P., & Partchev, I. (2016). Attempting to differentiate fast and slow intelligence: Using generalized item response trees to examine the role of speed on intelligence tests. *Intelligence*, *56*, 82–92.

Dodonov, Y. S., & Dodonova, Y. A. (2012). Response time analysis in cognitive tasks with increasing difficulty. *Intelligence*, *40*, 379–394.

Dorociak, K. E., Mattek, N., Lee, J., Leese, M. I., Bouranis, N., Imtiaz, D., Doane, B. M., Bernstein, J. P. K., Kaye, J. A., & Hughes, A. M. (2021). The survey for memory, attention and reaction time (SMART): Development and validation of a brief web-based measure of cognition for older adults. *Gerontologia*, *67*, 740–752.

Eizaguirre, M. D., S. I., Vanotti, Aguayo Arelis, A., Rabago Barajas, B., Cores, E. V., Macías, M. A., Benedict, R. H. B., & Cáceres, F. (2020). Symbol digit modalities test-oral version: An analysis of culture influence on a processing speed test in Argentina, Mexico, and the USA. *Developmental Neuropsychology*, *45*, 129–138.

Ganguli, M., Ratcliff, G., Chandra, V., Sharma, S., Gilby, J., Pandav, R., Belle, S., Ryan, C., Baker, C., Seaberg, E., & Dekosky, S. (1995). A hindi version of the MMSE: The development of a cognitive screening instrument for a largely illiterate rural elderly population in india. *International Journal of Geriatric Psychiatry*, *10*, 367–377.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*, 608–626.

Gross, A. L., Jones, R. N., Fong, T. G., Tommet, D., & Inouye, S. K. (2014). Calibration and validation of an innovative approach for estimating general cognitive performance. *Neuroepidemiology*, *42*, 144–153.

Gross, A. L., Khobragade, P. Y., Meijer, E., & Saxton, J. A. (2020). Measurement and structure of cognition in the longitudinal aging study in India-diagnostic assessment of Dementia. *Journal of the American Geriatrics Society*, *68*, S11–S19.

Hess, B. J., Johnston, M. M., & Lipner, R. S. (2013). The impact of item format and examinee characteristics on response times. *International Journal of Testing*, *13*, 295–313.

Humphreys, G. W., Duta, M. D., Montana, L., Demeyere, N., McCrory, C., Rohr, J., Kahn, K., Tollman, S., & Berkman, L. (2017). Cognitive function in low-income and low-literacy settings: Validation of the tablet-based oxford cognitive screen in the health and aging in Africa: A longitudinal study of an INDEPTH community in South Africa (HAALSI). *The Journals of Gerontology: Series B*, *72*, 38–50.

Junghaenel, D. U., Schneider, S., Orriens, B., Jin, H., Lee, P.-J., Kapteyn, A., Stone, A. A., Meijer, E., Zelinski, E. M., Hernandez, R., & Stone, A. A. (2023). Inferring cognitive abilities from response times to web-administered survey items in a population-representative sample. *Journal of Intelligence*, *11*, 3.

Koski, L., Xie, H., & Konsztowicz, S. (2011). Improving precision in the quantification of cognition using the Montreal cognitive assessment and the mini-mental state examination. *International Psychogeriatrics*, *23*, 1107–1115.

Kyllonen, P. C., & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence*, *4*, 14.

Langa, K M., Ryan, L H., McCammon, R J., Jones, R N., Manly, J J., Levine, D A., Sonnega, A., Farron, M., & Weir, D R. (2020). The health and retirement study harmonized cognitive assessment protocol project: Study design and methods. *Neuroepidemiology*, *54*, 64–74.

Lee, J., Banerjee, J., Khobragade, P. Y., Angrisani, M., & Dey, A. B. (2019). LASI-DAD study: A protocol for a prospective cohort study of late-life cognition and dementia in India. *BMJ Open*, *9*, e030300.

Lee, J., Ganguli, M., Weerman, A., Chien, S., Lee, D. Y., Varghese, M., & Dey, A. B. (2020). Online clinical consensus diagnosis of dementia: Development and validation. *Journal of the American Geriatrics Society*, *68*, S54–S59.

Lee, J., Khobragade, P. Y., Banerjee, J., Chien, S., Angrisani, M., Perianayagam, A., Bloom, D. E., & Dey, A. B. (2020). Design and methodology of the longitudinal aging study in India-diagnostic assessment of dementia (LASI-DAD). *Journal of the American Geriatrics Society*, *68*, S5–S10.

Levine, R. N. (2008). A geography of time: On tempo, culture, and the pace of life. Basic Books. ISBN: 0786722533. https://books.google.com/books?hl=en&lr=&id=IXq7CdAiQOQC&oi=fnd&pg=PR9&dq=a+geography+of+time&ots=SvMGKryJ6g&sig=tA4fC0M01XG1qkcfIzSSIqyv0dE.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, *103*, 403–428.

Sanders, S., Schofield, L. S., Schumm, L. P., & Waite, L. (2024). Measuring cognitive function and cognitive decline with response time data in NSHAP. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*.

Schmid, N. S., Taylor, K. I., Foldi, N. S., Berres, M., & Monsch, A. U. (2013). Neuropsychological signs of Alzheimer's Disease 8 Years prior to diagnosis. *Journal of Alzheimer's Disease*, *34*, 537–546.

Schneider, S., Junghaenel, D. U., Meijer, E., Stone, A. A., Orriens, B., Jin, H., Zelinski, E. M., Lee, P. J., Hernandez, R., & Kapteyn, A. (2023). Using item response times in online questionnaires to detect mild cognitive impairment.

The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences, *78*, 1278–1283.

Seelye, A., Mattek, N., Sharma, N., Riley, T., Austin, J., Wild, K., Dodge, H. H., Lore, E., & Kaye, J. (2018). Weekly observations of online survey metadata obtained through home computer use allow for detection of changes in everyday cognition before transition to MCI. *Alzheimer's & Dementia : The Journal of the Alzheimer's Association*, *14*, 187–194.

Shega, J. W., Sunkara, P. D., Kotwal, A., Kern, D. W., Henning, S. L., McClintock, M. K., Schumm, P., Waite, L. J., & Dale, W. (2014). Measuring cognition: The chicago cognitive function measure in the national social life, health and aging project, wave 2. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *69*, S166–S176.

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*, 5–20.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*, 323–339.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*, 67–85.

Wylie, S. A., Ridderinkhof, K. R., Eckerle, M. K., & Manning, C. A. (2007). Inefficient response inhibition in individuals with mild cognitive impairment. *Neuropsychologia*, *45*, 1408–1419.