Social Science Computer Review
1-15

DOI: 10.1177/0894439317746328
journals.sagepub.com/home/ssc
SAGE

The Effect of Online and Mixed-Mode Measurement of Cognitive Ability

Tarek Al Baghal[1]

Abstract

A number of studies, particularly longitudinal surveys, are collecting direct measures of cognitive ability, given its importance as a measure in social science research. As longitudinal studies increasingly switch to mixed-mode data collection, frequently including a web component, differences in survey outcomes including cognitive ability may result from mode effects. Differences may arise due to respondent self-selection into mode or due to the mode causing differential measurement. Using a longitudinal survey that measured cognitive ability after introducing a mixed-mode design with a web component, this research explores if and how mode affects cognitive ability outcomes. This survey allows for control of several possible selection mechanisms, including a limited set of direct cognitive ability measures collected in a single mode in an earlier wave. Findings presented here show clearly that web respondents do better on a number of cognitive ability indicators. However, it does not appear that this is wholly explainable by respondents of different ability self-selecting into particular modes. Rather, it appears that measurement of cognitive ability may differ across modes. This result is potentially problematic as comparability is a key component of using cognitive ability in further research.

Keywords

online surveys, mixed-mode surveys, cognitive ability, self-selection

Cognitive ability is an important construct for many studies that use survey data. It has been used to understand a number of outcomes in all aspects of human life (e.g., Denny & Doyle, 2008; Heckman, Stixrud, & Urzua, 2006; Nyborg, 2009) and survey data quality (e.g., Al Baghal, 2017; Couper, Tourangeau, Conrad, & Zhang, 2013; Struminskaya, 2016). Cognitive ability is also an important outcome of interest on its own to understand cognitive development and decline. Several longitudinal studies exploring aging or birth cohorts have collected direct measures of cognitive ability due to interest in cognitive development and decline. Some notable examples include the English

[1] Institute for Social and Economic Research, University of Essex, Colchester, United Kingdom

Corresponding Author:
Tarek Al Baghal, Institute for Social and Economic Research, University of Essex, Room 2N2.4.26, Wivenhow Park, Colchester CO4 3SQ, United Kingdom.
Email: talbag@essex.ac.uk

Longitudinal Study on Ageing (ELSA; Banks, Breeze, Lessof, & Nazroo, 2006), the U.S. Health and Retirement Study (HRS; Crimmins, Kim, Langa, & Weir, 2011); the Survey of Health and Retirement in Europe (Börsch-Supan et al., 2013), the National Study of Health and Development (NSHD; Kuh et al., 2011), and the 1958 British Birth Cohort (Power & Elliot, 2006).

In most general population studies, cognitive ability is rarely measured directly, with studies relying on proxies such as age and education. However, some general population household longitudinal studies have tested cognitive ability, reflecting the importance of accurate measures compared to proxies: Understanding Society: The United Kingdom Household Longitudinal Study (UKHLS; McFall, 2013); the Household, Income, and Labour Dynamics in Australia Survey (HILDA; Summerfield et al., 2016); and the German Socio-Economic Panel (SOEP; Wagner, Frick, & Schupp, 2007). While the number of surveys collecting these measuring is increasing, the recency means there is little research on measuring cognitive ability in a large general population survey. Given the increasing collection of cognitive ability measures in surveys, understanding how cognitive ability is measured is an important empirical question.

Due largely to cost considerations, longitudinal surveys are increasingly incorporating mixed-mode designs, whereby some respondents are interviewed using one mode, while others are interviewed with another (e.g., Jäckle, Burton, & Lynn, 2015). Web surveys are often chosen as part of a mixed-mode design for these practical and cost reasons (de Leeuw, 2005). However, a number of studies comparing web and interviewer-administered surveys show that significant response distribution differences can exist between modes (Duffy, Smith, Terhanian, & Bremer, 2005; Jäckle, Roberts, & Lynn, 2010; Lugtig, Lensvelt-Mulders, Frerichs, & Greven, 2011). How these different modes may affect measurement of cognitive ability in particular is not well understood, with few studies exploring mode effects in nationally represented studies. The possibility of mode differences in measuring cognitive ability within a mixed-mode design is of particular importance as standardized results are necessary for comparability (Lang, Weiss, Stocker, & von Rosenbladt, 2007).

Evidence on the Impact of Mode

The limited available evidence suggests the possibility that using different modes can lead to differences in cognitive ability measures. Both the HRS and HILDA found that telephone respondents scored higher on cognitive ability measures than face-to-face respondents (Herzog & Rodgers, 1997; Wooden, 2013). However, both findings could be explained by self-selection and design. HRS respondents aged 80 and older were all assigned to face-to-face surveys; those aged 70–79 were assigned to telephone but could switch modes based on preference. HILDA interviewed respondents by telephone only as a last resort when previous invitations to complete the survey were unsuccessful.

Although telephone respondents at the first wave of the HRS were younger, healthier, and more educated on average, attempts to account for this self-selection using statistical controls reduced but did not remove the difference in mode effect (Herzog & Rodgers, 1997). However, an experiment in the second and third waves of the HRS randomly assigned respondents aged 79–81 to complete either telephone or face-to-face versions only. Cognitive ability scores did not differ across modes among those randomly assigned, suggesting the initial difference arose from (incompletely controlled) self-selection (Herzog & Rodgers, 1999).

Studies comparing web to interviewer-administered modes suggest potential differences in outcomes. In one, members of a web panel answered more political knowledge questions correctly than telephone survey respondents (Strabac & Aalberg, 2011). Studies that randomly assigned respondents to modes have also found potential measurement differences. Fricker, Galesic, Tourangeau, and Yan (2005) randomly assigned respondents to either telephone or web modes, finding web respondents answered more knowledge questions correctly. Further, a recent experiment found

differences in cognitive ability measures across randomly assigned survey modes (Gooch, 2015). Respondents were assigned to complete an interviewer-administered or self-administered computer survey in controlled settings, using a short, four-word verbal test as a proxy for cognitive ability. The marginal distributions show mode differences for each question, with both modes having higher scores than the other for two questions each. Additional analyses using item response theory models suggest that the scale differentiates cognitive ability similarly across modes. Differences appear to arise because people react to the questions differently in the different modes.

There are several possible reasons for these differences to arise. First, presenting the survey either aurally or visually can directly affect the cognitive burden placed on the respondent (Tourangeau, Rips, & Rasinski, 2000). The presence of interviewers can also impact people's survey answers in a variety of ways and may also affect cognitive ability measures. For example, they may offer help or encouragement, lead to faster responses due to conversational norms, reduce satisficing tendencies, or may cause the respondent to feel pressure to appear knowledgeable (de Leeuw, 2005). Depending on the task, web respondents could also "cheat" by using assistance such as calculators or web search engines (e.g., Munzert & Selb, 2017). In household surveys, other members of the household could learn about the test from members participating earlier.

These factors suggest differences in measurement across modes; however, self-selection may lead to differences in sequential mixed-mode designs (Jäckle et al., 2015). Web respondents may differ from others in that they have the cognitive ability to use a computer, and older respondents technically inclined to use computers may differ from other elderly respondents who do not (Petchev, 2009). As such, where cognitive ability is found to be different across web and interviewer modes, it may be that differences are due to sample selection and not measurement.

The small number of studies on the possibility that mode affects cognitive ability outcomes leaves a number of questions unanswered. First, to date, it is not clear if mode differences occur in measuring cognitive ability in a representative general population study using comparable computerized self-administered modes. Second, if differences do occur, a better understanding is needed if these are due to self-selection or measurement. This research will examine possible differences in cognitive ability scores across web and interviewer-administered survey modes. This study takes advantage of a large representative longitudinal survey that introduced a sequential mixed-mode (web then face to face) design and measured cognitive ability with questions specifically to test numeric skills, inductive reasoning, and working memory. Unlike other studies comparing cognitive ability across these modes, the channel for presentation of questions is constant, that is, all are presented visually.

Data and Method

Sample

The Innovation Panel (IP) longitudinal survey is part of UKHLS. The IP is a vehicle for experimentation regarding aspects of survey design in a longitudinal survey context. It uses a multistage probability sample of persons and households in England, Scotland, and Wales. At the fourth wave (IP4) and seventh wave of IP (IP7), refreshment samples were also drawn. Waves are conducted annually, and interviews are attempted with all household members 16 years of age and older (full details at www.understandingsociety.ac.uk). The cognitive ability battery of interest was conducted as part of IP7. An initial smaller set of cognitive ability items was also asked at the third wave of IP (IP3), so only of the initial wave IP (IP1) respondents.

Prior to the fifth wave of IP (IP5), all interviews were conducted by interviewers. At IP5, a random two thirds of sample households were allocated to a sequential mixed-mode web and Computer Assisted Personal Interview (CAPI) design, while the other third were administered the

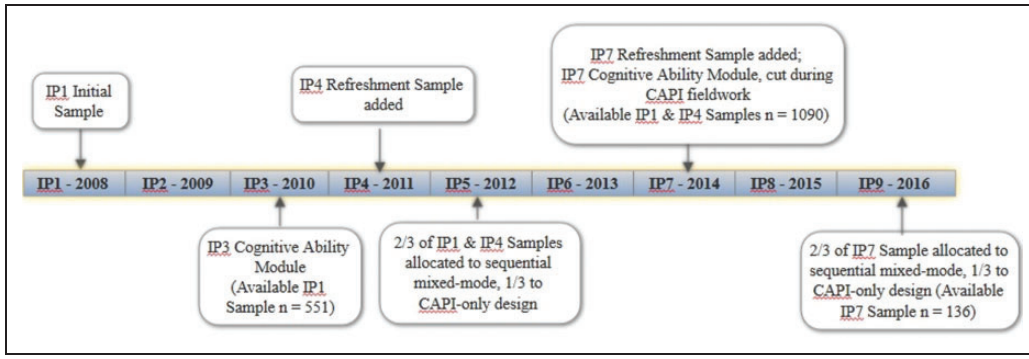

Figure 1. Time line of the Innovation Panel: Samples, mixed-mode, and cognitive ability measures.

standard single-mode CAPI design. In the mixed-mode treatment, if any household member did not respond to the web survey within 2 weeks, an interviewer was sent to attempt a face-to-face interview. This allocation remained the same for the original and IP4 refreshment samples through IP7. Respondents could access the web survey by tablet, but smartphone users were blocked, with a web page informing them to complete on another device. The majority of web respondents at IP7 completed via a PC ($n = 610$, 81.8%), with a minority of tablet respondents ($n = 137$, 18.2%). The IP7 refreshment was allocated to the mixed-mode design at the ninth wave of IP (IP9). Thus, the target cognitive ability measures are available prior to any possible self-selection. The time line for samples, introduction of mixed-mode designs, and measurements of cognitive ability are outlined in Figure 1.

Response rates for the IP are calculated as completion rates among those responding at their initial wave of interview. At IP1, conducted in 2008, the individual response rate by IP1 sample members was 52.4%. In 2011, for the IP4 refreshment sample, the initial response rate was 44.1%. The IP7 completion rate among IP1 respondents was 38.4%, producing a net response rate of 20.1% (AAPOR response rate 3 (RR3)). The IP7 completion rate among the IP4 refreshment sample was 62.1%, producing a net response rate of 27.4% (AAPOR RR3). For the IP7 refreshment sample, the individual response rate at IP7 sample was 24.3%. The IP9 completion rate for the IP7 refreshment sample was 82.7%, producing a net response rate of 20.1%.

Measuring Cognitive Ability

At IP7. For both CAPI and web respondents, the cognitive ability module was computerized. In face-to-face interviews, this module was administered as a computer-assisted self-interview (CASI), where the interviewer turned their laptop around for the respondent to complete on their own. The visual design of the survey and ordering of questions were exactly the same for web and CASI versions, and so this possible factor should be controlled. The visual design for the tablet was not optimized for the tablet and the same as displayed on a PC. There was no effect on the total of correctly answered questions by device among web respondents, $F(1, 669) = .28$, $p = .59$, and will not be considered separately.

Cognitive ability measures are largely based on those used in the HRS and ELSA to measure executive reasoning and working memory. Each type of question was prefaced with an explanation of what the question was asking (see Online Appendix A for full questions). Executive reasoning measures began with a set of three number series questions were asked as in the HRS (Fisher, McArdle, McCammon, Sonnega, & Weir, 2013). Number series questions measure quantitative reasoning, using concepts depending on mathematical relationships. Three questions of verbal

analogies followed, designed as in the HRS (Fisher et al., 2013), in order to measure verbal reasoning. Still, the majority (52%) chose the incorrect analogy of "town." The module ended with two numeracy questions, taken from the HRS (Ofstedal, Fisher, & Herzog, 2005) and ELSA (Banks et al., 2006). All of the questions with a unique correct answer testing inductive reasoning, that is, number series, verbal analogies, and numeracy questions are scored as 1 if answered correctly, 0 if answered incorrectly. "Don't know" responses were considered incorrect and scored as 0.

Prior to the final two numeracy questions, a series of four forward digit-span tasks were given, frequently used as a working memory test (Richardson, 2007). In each, a number string of differing lengths was presented for 1 s for each number in the number string. For example, a three-digit string would be presented on the screen for 3 s, while a seven-digit string would appear for 7 s. In order of occurrence, number lengths of three, seven, nine, and five digits were asked. A respondent's digit-span working memory is commonly indicated by the longest digit span they got entirely correct (Conway et al., 2005). The number length is the best of the four trials or 0 if the respondent did not get any fully correct. There is uncertainty in meaning of don't know responses to digit-span tasks, in that a respondent could not remember any of the digits or only some or the wrong ordering. As such scoring all don't know responses as meaning the respondent have a zero digit-span working memory seems inappropriate. These small number of respondents who answered don't know on every digit-span trial ($n = 7$) are therefore not included in the analyses of digit spans.

During fieldwork, it was determined the survey was running too long and would cost more than the budget allowed. In response, a number of modules were dropped from the survey, including the cognitive ability module. There was no reduction in the web version, where fielding took place earlier and the additional questions did not add to interviewing cost. A total of 338 IP1 or IP4 sample respondents completed the module in CAPI before the module was dropped. In all, more than half of all IP1/IP4 respondents had already completed the module; a total of 1,090 (752 via web, 338 via CAPI) completed the cognitive ability questions. Of these completing respondents, 1,083 completed at least one of the digit-span tasks, with varying levels in the number of don't know responses for each separate task ranging from 15 to 36. For the IP7 refreshment sample, 211 respondents completed the cognitive ability module at IP7 before it was dropped; of these, 136 also took part in IP9, where 31 completed the CAPI-only version, 36 completed the mixed-mode CAPI, and 69 completed via the web.

Fieldwork is not prioritized on respondent characteristics by the fieldwork agency. Those completing the cognitive ability module are therefore likely to be similar in important characteristics to those who did not. Tests comparing respondent demographics, for example, age, race, and education, show no significant differences, providing evidence to this claim. As a further test, for the IP1 sample, IP3 cognitive ability scores (outlined below) were compared between those completing the IP7 questions in CAPI (since all in web were asked) and those respondents in CAPI not asked the cognitive ability module. There is no difference in prospective memory, $F(1,500) = .83$, $p = .36$; the serial seven subtraction task, $F(1,471) = .43$, $p = .52$; or the FAS test, $F(1,490) = .11$, $p = .74$. Taken together, this is suggestive that those participating earlier during CAPI fieldwork (and hence asked the cognitive ability questions) are similar to those interviewed later.

However, to account for the differences in sample size, the data were weighted to the final completed sample by mode condition. All web respondents completed the cognitive ability module. However, 115 mixed-mode CAPI respondents answered these questions of a total 369 mixed-mode CAPI respondents interviewed. Similarly, 223 of the 528 CAPI-only interviews include the cognitive ability module. These numbers were weighted to the end sample achieved by taking the inverse of the proportion that completed the cognitive ability module. Since all completed the cognitive ability module, web respondents are given a weight of 1. Mixed-mode CAPI respondents are given a weight of $1/(369/115) = 3.209$, and CAPI-only a weight of $1/(528/223) = 2.368$. All analyses reported are weighted; however, all analyses were also conducted without weights for sensitivity tests and did not find substantive differences.

At IP3. A smaller number of different cognitive ability measures were also asked at IP3 in preparation for inclusion in the main UKHLS survey. Three measures were used: a prospective memory task, a phonemic fluency task (FAS test), and the serial seven subtraction task (see Online Appendix A). At the beginning of the cognitive ability questions, respondents were informed that at some point during the interview, the interviewer would hand them a pen and paper and asked to write their date of birth on the upper-left corner of the paper. For all respondents, the interviewer handed them the pen and paper following the next task for phonemic fluency. This prospective memory measure has limited variability. Among the IP3 respondents who also completed the IP7 cognitive ability module ($n = 551$), 92.2% correctly wrote their date of the birth on the upper left of the paper. For the remainder, 4.2% wrote their date of birth, but not in the left-hand corner, 1.6% wrote something other than their birthday in the upper-left corner, while 2.0% did something else.

The mean number of words recalled in this FAS test among those also answering the IP7 cognitive ability module is 13.16. The FAS is highly skewed to the right (skewness = .352), so the natural log is taken. A small value (.00001) is added to 0 FAS scores to take the log, affecting a small number of cases ($n = 24$, 1.5%). Sensitivity tests with smaller values showed no differences in outcome. After next concluding the prospective memory task, respondents were then asked to subtract 7 from 100 and continue subtracting 7 from each subsequent answer for a total of five subtractions. Interviewers recorded the number of correct answers from 0 to 5. Most respondents who also responded to the Wave 7 cognitive ability module successfully made all subtractions correctly: 72.0% correctly subtracted all five attempts, 13.2% four correct, 10.0% three correct, 3.0% two correct, 1.5% one correct, and 0.4% none correct. The mean number correct was 4.5 ($SD = 0.94$).

Analysis Methods

Only IP1 and IP4 sample members are included in the analyses exploring mode difference at IP7 for two reasons. First, in their initial year, the IP7 refreshment samples were all surveyed in person. Second, some of the important explanatory measures were collected at prior waves and are thus not available for these new respondents. To account for possible mode selection effects in the mixed-mode survey, significant correlates with response in the mixed-mode design identified by Jäckle, Burton, and Lynn (2015) are included as covariates in the following analyses. The only variables that were jointly significant in predicting individual response rates are urban location and respondents saying there was definitely no chance of responding to a web survey. However, web use and provision of an e-mail address were also related to whether everyone in the household completed the survey or not. The likelihood of responding to a web survey was initially asked at IP4 before the switch to a mixed-mode design. Those saying there is "no chance" of responding by web are coded as 1, all else are coded as 0. Internet use is indicated as a dummy variable, with 1 for those saying they used the Internet every day, 0 for all else. Those who ever provided an e-mail contact were coded equal to 1, with those who had not coded equal to 0.

Finally, as a further indicator for economic status, monthly personal income (in Great Britain pounds) is included as additional control for mode selection, as it provides the resources for the requisite devices to complete a web survey. Those with greater cognitive ability may also be expected to earn more. The released data include imputed data for respondents with missing income data, imputing using a number of techniques (Knies, 2016). The imputed data are included in the following analyses, so there are no missing values for income. Characteristics of these respondent measures for those completing the cognitive ability module by mode design allocation and mode of completion are presented in Table 1. Different superscript alphabets in all following tables indicate statistically significant differences across mode conditions. Outcomes with the same superscript within a row indicate no significant difference across mode conditions.

Table 1. Respondent Characteristics by the Mode of Completion.

Variables	Mean—CAPI Only (*SD*; *n*)	Mean—MM CAPI (*SD*; *n*)	Mean—MM Web (*SD*; *n*)
Age	52.87[a] (27.41; 528)	62.63[b] (32.34; 369)	47.63[c] (17.07; 752)
Income	1,597.73[a] (2,288.84; 528)	1,289.29[b] (1,615.18; 369)	1,782.90[a] (1,518.85; 752)
College/professional degree	0.22[a] (0.63; 528)	0.08[b] (0.48; 369)	0.31[c] (0.46; 752)
Urban	0.76[a] (0.66; 528)	0.61[b] (0.88; 369)	0.74[a] (0.44; 752)
Understand questionnaire	0.73[a] (0.68; 528)	0.57[b] (0.89; 369)	0.72[a] (0.45; 657)
Unlikely to respond by web	0.30[a] (0.71; 459.34)	0.60[b] (0.88; 327.29)	0.15[c] (0.35; 632)
Daily Internet use	0.66[a] (0.73; 525.63)	0.43[b] (0.89; 369)	0.74[c] (0.44; 752)
Provided e-mail address	0.24[a] (0.66; 528)	0.15[b] (0.64; 369)	0.29[c] (0.45; 752)

Note. Weighted *n* shown. Superscripts a, b, and c within row indicate significant difference between mode conditions at $p < .05$.
CAPI = Computer Assisted Personal Interview, MM = Mixed-Mode.

There are several differences across those responding in the different modes that suggest how selection could affect cognitive ability measures. Those responding via the web are significantly younger, more educated, report daily Internet use more, and are more likely to provide an e-mail address than CAPI-only or those in the mixed-mode condition responding via CAPI. Those responding via the web also report less that they are unlikely to respond via the web than those responding to either CAPI group, with significantly more mixed-mode CAPI respondents reporting this unlikelihood than CAPI-only respondents.

Web and CAPI-only respondents are not significantly different in terms of income, but both earn significantly more than mixed-mode CAPI respondents. Similarly, there is no difference in CAPI-only and web respondents for the proportion rated having excellent questionnaire understanding, but both are greater proportions than the mixed-mode CAPI respondents. CAPI-only respondents have more college graduates, report more using the Internet daily, and are more likely to provide an e-mail address than mixed-mode CAPI respondents, being between these and web respondents on these variables. Additionally, significantly less mixed-mode CAPI respondents live in urban areas than those in either other mode group.

To initially explore and disentangle selection and measurement effects in mixed-mode designs, propensity score methods are used (Lugtig et al., 2011). Propensity scores in this study estimate the conditional probability to be a respondent in the CAPI-only, mixed-mode CAPI, or mixed-mode web design. Propensity scores are estimated by logistic regression models using the respondent characteristics potentially related to mode selection as covariates in Table 1. Inverse probability of treatment weighting (IPTW) methods are utilized, using weights based on the propensity score where the selection-related covariates are independent of treatment (i.e., mode; Schonlau et al., 2004). Using these weights allow for estimation of the average treatment effect for multiple conditions.

Multivariate models are then used to estimate the impact of mode at IP7 while controlling for other possible indicators of selection or measurement effects. Multilevel logistic regressions estimate the effects of selected variables on correct responses to the eight inductive reasoning and numeracy questions. These models account for the dichotomous nature of the outcome variable (correct or incorrect response) as well as the structure of the data as responses are nested within respondents. Random intercept models are used, with the one random effect occurring at the respondents level. To model digit span, Poisson regression models are used, given the count nature of the response. Negative binomial models were also tested, but the dispersion parameters were small and not different from 0, and the coefficients using the different distributions were generally indistinguishable.

The multilevel models include fixed effects for each of the specific questions to control for different question difficulty. Both multilevel and Poisson regression models include indicators of

response time for the respective questions, captured from survey paradata, to account for the amount of cognitive effort used (e.g., Yan & Tourangeau, 2008). For the IP1 sample, cognitive ability measured at the third wave can be used as a more direct control to identify possible selection differences on the outcome of interest. For both samples, age at the time of the survey and education are included as proxies for cognitive ability (Schwarz & Knäuper, 1999). Age squared is also included to indicate possible changes over the life course. Education is coded as those having a university or professional degree or not.

An additional proxy measure for cognitive ability comes from the subjective rating of the interviewer if the respondent understood the questionnaire or not on a 5-point scale completed after the interview. Most respondents' understanding of the questionnaire was deemed "excellent" at every wave, and questionnaire understanding is indicated as excellent versus any worse rating of understanding. To avoid the effect that asking cognitive ability questions may have on this measure at IP7, the interviewer's report of understanding is based on whatever is the most recent measurement prior to IP7, given the instrument remains generally the same over time. Interviewer ratings of understanding taken at IP7 are used for those people responding for the first time and have no other measure to avoid loss of cases. Some web respondents do not have any rating due to the interviewer not responding when they had responded via CAPI in earlier waves. There are 95 of these missing cases not included in the multivariate models.

Additional tests of cognitive ability on subsets of the total sample provide additional insight into the possibilities of self-selection or measurement differences. For IP1 respondents, the IP3 cognitive test provides comparative measures in a single mode prior to any possible self-selection. Similarly, for IP7 refreshment sample members, the cognitive ability measures in question were all asked in a CAPI design prior to the introduction of mixed-modes for these sample members at IP9. The possibility for measurement differences is bolstered if those self-selecting into the web survey for the first time at IP9 do not differ on these target measures taken at IP7 from those IP9 CAPI respondents. Conversely, the possibility of self-selection differences is further suggested if IP7 refreshment samples selecting into web at IP9 are significantly different in cognitive ability measures taken at IP7. However, due to the dropping of the cognitive ability module at IP7 in the CAPI design (which all IP7 refreshment sample members took part in) and attrition, the number of cases for this test is reduced.

Results

The outcome to the inductive reasoning and numeracy cognitive ability questions by each of the mode conditions are presented in Table 2. Strikingly, web respondents answered correctly significantly more often than either CAPI condition for every measure leading to significantly more correct responses overall. Nearly, all of web respondents answered several of the questions correctly, and more than half (50.6%) answered correctly even for the most frequently missed question, the third verbal analogy. Web respondents answered 6.91 of the 8 inductive reasoning and numeracy questions correctly on average, with 84.3% of all the questions being answered correctly via the web. Generally, the number correct for the CAPI-only condition falls between that of the web group on the high end and the mixed-mode CAPI respondents on the low end. This intermediate level is seen in the percentage correct in five of the eight questions (Number Series 2, all three verbal analogies, and used book cost). For these, CAPI-only respondents on the whole did significantly better than mixed-mode CAPI respondent and worse than web respondents. These differences are reflected in a similar pattern for the total number of respondents who correctly answered.

The propensity score weighted (IPTW) means for total questions correct present results further suggesting the possibility of measurement differences. By conditioning weights on possible selection-related covariates found in previous research, the IPTW means attempt to provide an

Table 2. Cognitive Ability Measures and Propensity Score Weighted Mean by Mode Condition.

Measure	% Correct—CAPI Only	% Correct—MM CAPI	% Correct—MM Web
Number Series 1	86.2[a]	84.4[a]	96.8[b]
Number Series 2	75.5[a]	66.1[b]	87.4[c]
Number Series 3	91.5[a]	90.4[a]	97.3[b]
Verbal Analogy 1	93.3[a]	88.7[b]	97.3[c]
Verbal Analogy 2	75.9[a]	58.3[b]	85.8[c]
Verbal Analogy 3	38.4[a]	31.3[b]	50.6[c]
Used book cost	61.0[a]	53.9[b]	80.2[c]
Disease rate	87.0[a]	83.5[a]	96.0[b]
Mean total correct	6.08[a]	5.57[b]	6.91[c]
IPTW mean	6.13[a]	6.04[a]	6.76[b]
Digit-Span Memory	**% Best Recalled—CAPI Only**	**% Best Recalled—MM CAPI**	**% Best Recalled—MM Web**
0	8.6	11.7	0.5
3	9.6	7.2	3.9
5	30.0	38.7	17.8
7	37.7	34.2	41.5
9	14.1	8.1	36.3

Note. Superscripts a, b, and c within row indicate significant difference between mode conditions at $p < .05$. $\chi^2_8 = 240.94$, $p < .0001$.
CAPI = Computer Assisted Personal Interview, MM = Mixed-Mode.

estimate if self-selection was limited. The IPTW adjustment brings each mean closer in value to the others than unadjusted numbers, with slight increases in value for CAPI-only and mixed-mode CAPI, and a decreased total correct mean in web. Importantly, the differences between these, the estimated average treatment effect, are not significant between CAPI versions unlike the unweighted data. However, the average treatment effect leads to a significant larger IPTW web mean than either CAPI condition, suggesting there may be measurement differences between modes.

A similar pattern is seen in the forward digit-span memory task shown in bottom part of Table 2. Web respondents had significantly higher digit spans than either of the CAPI conditions with CAPI-only respondents doing worse than those in the web but better than those in the mixed-mode CAPI condition. More than a third of web respondents could recall the longest digit span given, a percentage more than 2.5 times that of CAPI-only respondents and nearly 4.5 times more than the mixed-mode CAPI condition. Conversely, relatively many times more of both CAPI conditions failed to complete any of the digit spans fully than did web respondents.

The results above show that the observed outcomes of the cognitive ability measures differ by mode. It appears possible that the CAPI-only group is a mixture of people who would self-select to be web respondents or CAPI respondents like the mixed-mode condition, given the choice. If the web respondents actually have higher cognitive ability than the mixed-mode CAPI respondents as results suggest, the CAPI-only results should be in between these in the fashion of a weighted average. This intermediate result is generally observed for inductive reasoning, numeracy, and digit recall questions. Evidence suggesting against selection effects is also found in these tables, however. While in every instance web respondents provide a significantly greater percentage of correct responses, this is not true of the CAPI-only comparison to the mixed-mode CAPI group. The three instances of nonsignificant differences at least suggest more similarity between the CAPI groups than either to the web respondents.

To further explore this possibility, several multivariate models are estimated with the goal of controlling a number of factors to separate and identify either selection or measurement effects.

Table 3. Multivariate Models Predicting Correct Response and Forward Digit Span.

Indicator	Odds Ratios for Correct Response, Random and Fixed Effects		Poisson Regression Coefficients, Forward Digit Span	
	IP1 and IP4 Sample	IP1-Only Sample	IP1 and IP4 Sample	IP1-Only Sample
Mode condition (Reference: Web)				
CAPI Only	0.483*	0.480*	−0.192*	−0.274*
MM-CAPI	0.393*	0.430*	−0.214*	−0.242*
Response time (logged)	0.764*	0.793*	−0.008	0.258*
College/professional degree	1.644*	1.195	0.047	−0.019
Understand questionnaire	1.676*	1.456*	0.083*	0.132*
Age	1.020	1.004	0.007	−0.003
Age^2	1.000	1.000	−0.0001	−0.0001
Income	1.0001*	1.000	0.000007	−0.00001
Urban	0.802	0.891	−0.033	−0.035
Unlikely to respond by web	0.622*	0.504*	−0.121*	−0.154*
Provided e-mail address	0.977	0.850	0.031	0.052
Daily Internet use	1.222	1.117	0.031	0.018
IP1 sample member	0.898		−0.011	
Serial 7 score		1.546*		0.062*
FAS score (logged)		1.319*		0.193*
Prospective memory test		1.077		0.109
Question (Reference: Verbal Analogy 1)				
Verbal Analogy 2	0.117*	0.103*		
Verbal Analogy 3	0.016*	0.014*		
Number Series 1	0.498*	0.401*		
Number Series 2	0.156*	0.132*		
Number Series 3	0.790	0.931		
Used book cost	0.074*	0.062*		
Disease rate	0.444*	0.423*		
Constant			1.800*	0.494*
Random effects parameters				
Respondent variance	1.412	1.149		
ICC	0.300	0.258		
Number of responses	6,920	3,888		
Respondents	865	486	873	472

Note. ICC = Intraclass correlation, IP = Innovation Panel.
*$p < .05$.
CAPI = Computer Assisted Personal Interview, MM = Mixed-Mode.

First, multilevel logistic regression models examine the results to the inductive reasoning and numeracy questions, with each of the eight responses coded as correct or not nested with respondents. Two such models are estimated; the first using responses from both IP1 and IP4 sample members and the second with only those IP1 sample members who had cognitive ability measured at the third wave. Mode effects are estimated by comparison of CAPI-only and mixed-mode CAPI to the baseline of web respondents. Other controls included for possible selection or measurement effects are fixed effects for each question, response time, and respondent characteristics presented in Table 1. Poisson regression models estimate forward digit span using the same independent variables (except fixed effects for questions). Table 3 presents the results to these models.

These models suggest that after controlling for these other factors, mode effects still remain. In particular, respondents in both CAPI conditions are significantly less likely to answer an item correctly and give shorter number spans than those in the web condition. That web respondents are more likely to answer correctly mirrors findings in Table 3 but is now inclusive of controls for the question (for the multilevel models), the amount of time responding, proxies for cognitive ability, and measures identified as being related to selection. Further, evidence of possible difference between CAPI conditions disappears after adding these controls. In all of the models, the difference between the coefficients for CAPI-only and mixed-mode CAPI are not significantly different at the $p < .05$ level (not shown).

The inclusion of an indicator for sample membership suggests that there are no differences in outcomes between IP1 and IP4 samples. These nonsignificant estimates are important as these suggest against sample effects, and models including the three cognitive ability measure captured at the third wave as controls for the IP1 sample only are also presented. The effects of mode still persist after including these more direct measures of cognitive ability. As with the models including both samples, web respondents are estimated to be significantly more likely to provide a correct response and longer number series than either CAPI condition. The effects of the Serial 7 and FAS score are both significant and in the expected direction. Higher scores on these two cognitive ability measures are related to greater odds of providing a correct response. Further, greater Serial 7 and FAS scores are significantly related to longer digit spans as expected. Prospective memory is not significant related to provide a correct response or longer digit span (although the estimate is in the expected direction), possibly reflecting the lack of discrimination this measure provided noted above. Education is only significant in the correct responses model for both samples, but becomes nonsignificant when including the IP3 cognitive ability measures, possibly indicating these do capture latent cognitive ability. Those rated to understand the questionnaire excellently in prior waves are estimated to have significantly higher odds of providing a correct response and providing longer number series in all models.

While these models attempt to control for selection after measurement, some data exist capturing measurement of cognitive ability prior to selection, limiting this effect. For IP1 respondents, a limited set of cognitive ability measures were captured all in the same mode at IP3 prior to any self-selection into mode. For the IP7 refreshment sample, some respondents took the target cognitive ability measures under consideration in this research, all in the same mode and prior to any self-selection introduced only at IP9. There are some limitations to the analyses noted, including the limited measures available for the IP1 sample and the smaller numbers of the IP7 refreshment sample taking the cognitive ability module due to its being cut during fieldwork. However, if these measures differ across conditions in a similar way when measured in the same mode, it would provide stronger evidence for selection. If there are no differences, particularly when using the same IP7 measures, it may suggest that there are measurement effects due to mode. Outcomes for those respondents taking these cognitive ability measures before any mode self-selection could take place are presented in Table 4.

Examining the measures taken at the third wave for the IP1 sample first, the scores are generally similar across respondents by their mode of response at IP7 and there are no statistically significant differences. Directionally, each mode condition did better than the other two on exactly one of the measures each. Importantly, these measures correlate significantly more strongly with the total correct score at IP7 when the mode remains constant, that is, CAPI. For CAPI respondents, the correlation between total correct items (of the inductive reasoning and numeracy questions) and the Serial 7 score is 0.46, with the FAS score = 0.36, and with prospective memory = 0.16. While not perfect correlations, all are nonzero and in some cases larger than generally seen in social research. Comparatively, among the correlation between the web respondents' total correct score and the Serial 7 score is 0.28, the FAS score is 0.13, and prospective memory is 0.03. All of these are substantially smaller than those among CAPI respondents, with the prospective memory measure not significantly different from 0. That the correlations show similarity in cognitive ability outcomes

Table 4. IP3 Cognitive Ability Measures by IP7 Mode Condition.

Cognitive Ability Test	CAPI Only	MM CAPI	MM Web
IP3 measures—IP1 sample			
Serial 7 score	4.57	4.32	4.51
FAS score (logged)	2.29	2.12	2.41
Prospective memory	0.91	0.94	0.92
Base *n*	112	70	371
IP7 measures—IP7 sample			
Mean total correct	5.97	6.14	6.32
Mean longest number series	6.03	6.35	6.35
Base *n*	31	36	69

Note. IP = Innovation Panel.
CAPI = Computer Assisted Personal Interview, MM = Mixed-Mode.

when mode is constant but less so when mode has changed suggests the possibility that the mode of response is affecting the measurements.

Importantly, no differences are found for the IP7 refreshment sample on any of the target ability measures across the modes when selection was possible at a later wave. Given this similarity across all measures, only central tendencies are presented. Although web respondents directionally have somewhat larger mean correct, unlike for the larger IP1 and IP4 samples, mixed-mode CAPI and web respondents are more similar in outcomes than with the CAPI-only respondents. In this instance, the CAPI-only respondents do not appear to fall in between the two mixed-mode groups. It does not appear these respondents are not approximating a weighted mean of those who would self-select to be web respondents or CAPI respondents like the mixed-mode condition, given the choice. Although both these additional tests of cognitive ability measures taken prior to any self-selection are somewhat limited, these do provide important additional evidence to the multiple other analyses conducted above.

Discussion and Conclusions

The current study uses longitudinal data with a mixed-mode design experiment and cognitive ability measures, presenting a number of findings informative to research designs. Initial results show cognitive ability scores are significantly different across modes. Web respondents perform significantly better than CAPI respondents in both the measures of inductive reasoning and numeracy (with correct or incorrect responses) and recall (measured by forward digit span). The study maintained a consistent visual presentation across all modes, so differences are not attributable to aural versus visual. The initial analyses between CAPI conditions are less clear. The CAPI-only condition, where no selection is possible, at times has aggregated results between the mixed-mode conditions, possibly representing a combination of respondents who would select into either web of CAPI modes, given the chance. Importantly, however, using propensity score methods, weighted results suggest that the web version may lead to more correct responses, while CAPI conditions may be similar in outcomes.

Analysis of multivariate models as well as cognitive ability measures taken before any self-selection could occur, including using the target cognitive ability questions, suggest that differences identified by mode are not wholly attributable to selection. Rather, there appears to be differences in the measurement of cognitive ability by mode. Using controls based on research about mode self-selections using the same IP survey (in prior waves) and direct measures of cognitive ability, the impact of mode remains. The inclusion of these controls in web respondents has both higher likelihood of providing a correct response and longer digit spans. However, the differences between CAPI conditions are no longer significant. This similarity in CAPI conditions along with the

differences of these to the web condition after controls implies the possibility for mode differences in measurement. Measures of cognitive ability captured before any self-selection also show little difference between respondents that later self-select into the web or CAPI modes. This latter result suggests that it is not necessarily cognitive ability of respondents affecting self-selection but potentially the mode of measurement affecting outcomes.

This possibility is problematic as measurement differences due to mode mean incomparable indicators of the latent construct of cognitive ability, both across and within respondents. It would be difficult to interpret findings across respondents answering in differing modes in a cross section. In longitudinal settings, this problem would not only exist within a wave but also in potentially identifying change within respondents who have changed response mode across waves. It is therefore suggested that where possible to measure cognitive ability in a single mode.

This potential problem is suggested when exploring the effect of cognitive ability on attitudes on climate change measured in the IP, an attitude shown to be sensitive to cognitive ability (Fischer & Glenk, 2011). In the IP, respondents were asked whether they believed the United Kingdom will be affected by climate change in the next 30 years, with yes and no response options. Logistic regression models predicting this attitude using total number of the executive function correct, while controlling for mode of response and the respondent characteristics, are shown in Online Appendix B. The first model assumes the effect of the cognitive ability measures is the same for respondents, controlling for other factors. The second model allows for interactions between mode of response and cognitive ability, exploring how differences in outcomes across modes affect estimates. While the first model shows no effect of cognitive ability on the climate change attitude, the second model shows that cognitive ability affects the estimated probability for a given response, but only for web respondents. In this instance, all else held equal, those with higher cognitive ability in the web are less likely to answer that climate change will affect the United Kingdom in the next 30 years. Not accounting for this interaction between mode and cognitive ability, outcomes would miss this potentially important result.

The results of this study have some limitations to be addressed in further research. First, it is not apparent as to why these differences arise, particularly since all respondents used a computerized self-administration with the same visual design to answer. Several possibilities are outlined above, but it seems the presence of an interviewer has an impact in some way. However, it does not appear to be specific to individual interviewers; the for interviewers on the total correct score is 0.08, which is relative low. Additional research should also explore the differences of a CAPI environment not using CASI with these other modes.

The second limitation is the possibility that the data have not appropriately controlled for possible selection effects. While a number of variables that would be related to selection effects for these measures are used, including past measures of cognitive ability, there still may be other important ones missed. This possibility is suggested by the relatively high respondent ICC estimated for the correct response models. These suggest that some respondent characteristics important to cognitive ability outcomes are not captured. While it is not necessary these are related to selection, the possibility cannot be eliminated. Further, while measures of cognitive ability taken before self-selection exist, there are some issues in measurement at IP3, with small numbers available when using the target cognitive ability questions at IP9.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The author is funded by a research award from the UK Economic and Social

Research Council (award no. ES/K005146/1) for "Understanding Society: The UK Household Longitudinal Study, Phase 3."

Supplemental Material

Supplementary material for this article is available online.

References

Al Baghal, T. (2017). "Last year your answer was . . . ": The impact of dependent interviewing question wording on measures of change. *Field Methods*, *29*, 61–78.

Banks, J., Breeze, E., Lessof, C., & Nazroo, J. (2006). *Retirement, health and relationships of the older person in England: The English Longitudinal Study on Ageing* (Wave 2). London, England: Institute for Fiscal Studies.

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., . . . Zuber, S. (2013). Data resource profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology*, *42*, 992–1001.

Conway, A. R. A., Kane, M. J., Bunting, F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*, 769–786.

Couper, M. P., Tourangeau, R., Conrad, F. G., & Zhang, C. (2013). The design of grids in web surveys. *Social Science Computer Review*, *31*, 322–345.

Crimmins, E. M., Kim, J. K., Langa, K. M., & Weir, D. R. (2011). Assessment of cognition using surveys and neuropsychological assessment: The Health and Retirement Study and the Aging, Demographics, and Memory Study. *Journals of Gerontology: Psychological Sciences*, *66B*, i162–i171.

de Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics*, *21*, 233–255.

Denny, K., & Doyle, O. (2008). Political interest, cognitive ability and personality: Determinants of voter turnout in Britain. *British Journal of Political Science*, *38*, 291–310.

Duffy, B., Smith, K., Terhanian, G., & Bremer, G. (2005). Comparing data from online and face-to-face surveys. *International Journal of Market Research*, *47*, 615–639.

Fischer, A., & Glenk, K. (2011). One model fits all?—On the moderating role of emotional engagement and confusion in the elicitation of preferences for climate change adaptation policies. *Ecological Economics*, *70*, 1178–1188.

Fisher, G. G., McArdle, J. J., McCammon, R. J., Sonnega, A., & Weir, D. R. (2013). *New measures of fluid intelligence in the HRS* (HRS Documentation Report DR-027). Ann Arbor: Survey Research Center, University of Michigan.

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

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

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, *24*, 411–482.

Herzog, A. R., & Rodgers, W. L. (1997). Measures of cognitive functioning in the AHEAD study. *The Journals of Gerontology Series B*, *52B*, 37–48.

Herzog, A. R., & Rodgers, W. L. (1999). Cognitive performance measures in survey research on older adults. In N. Schwarz, D. C. Park, B. Knäuper, & S. Sudman (Eds.), *Aging, cognition, and self-reports*. (pp. 327–340) Philadelphia, PA: Psychology Press.

Jäckle, A., Lynn, P., & Burton, J. (2015). Going online with a face-to-face household panel: Effects of a mixed mode design on costs, participation rates and data quality. *Survey Research Methods*, *9*, 57–70.

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

Knies, G. (Ed.). (2016). *Understanding society—UK household longitudinal study: Wave 1–6, 2009–2015, user manual*. Colchester, England: University of Essex.

Kuh, D., Pierce, M., Adams, J., Deanfield, J., Ekelund, U., Friberg, P., ... Hardy, R. (2011). Cohort profile: Updating the cohort profile for the MRC National Survey of Health and Development: A new clinic-based data collection for ageing research. *International Journal of Epidemiology*, *40*, e1–e9.

Lang, F., Weiss, D., Stocker, A., & von Rosenbladt, B. (2007). Assessing cognitive capacities in computer-assisted survey research: Two ultra-short tests of intellectual ability in the German Socio-Economic Panel (SOEP). *Schmollers Jarbuch*, *127*, 183–192.

Lugtig, P., Lensvelt-Mulders, G. J. L. M., Frerichs, R., & Greven, A. (2011). Estimating nonresponse bias and mode effects in a mixed-mode survey. *International Journal of Market Research*, *53*, 669–686.

McFall, S. (2013). *Understanding society: UK household longitudinal study: Cognitive ability measures*. Colchester, England: University of Essex.

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

Nyborg, H. (2009). The intelligence–religiosity nexus: A representative study of white adolescent Americans. *Intelligence*, *37*, 81–93.

Ofstedal, M. B., Fisher, G. G., & Herzog, A. R. (2005). *Documentation of cognitive functioning measures in the Health and Retirement Study* (HRS Documentation Report DR-006). Ann Arbor: Survey Research Center, University of Michigan.

Petchev, A. (2009). Survey breakoff. *Public Opinion Quarterly*, *73*, 74–97.

Power, C., & Elliot, J. (2006). Cohort profile: 1958 British birth cohort (National Child Development Study). *International Journal of Epidemiology*, *35*, 34–41.

Richardson, J. T. (2007). Measures of short-term memory: A historical review. *Cortex*, *43*, 635–650.

Schonlau, M., Zpart, K., Simon, L. P., Sanstad, K. H., Marcus, S. M., Adams, J., ... Berry, S. H. (2004). A comparison between responses from a propensity-weighted web survey and an identical RDD survey. *Social Science Computer Review*, *22*, 128–138.

Schwarz, N., & Knäuper, B. (1999). Cognition, aging, and self-reports. In D. Park & N. Schwarz (Eds.), *Cognitive aging—A primer* (pp. 233–252). Philadelphia, PA: Psychology Press.

Strabac, Z., & Aalberg, T. (2011). Measuring political knowledge in telephone and web surveys: A cross-national a comparison. *Social Science Computer Review*, *29*, 175–192.

Struminskaya, B. (2016). Respondent conditioning in online panel surveys: Results of two field experiments. *Social Science Computer Review*, *34*, 95–115.

Summerfield, M., Freiden, S., Hahn, M., La, N., Ning, L., Macalalad, N., ... Wooden, M. (2016). *HILDA user manual—Release 15*. Melbourne, Australia: Melbourne Institute of Applied Economic and Social Research.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, MA: Cambridge University Press.

Wagner, G. G., Frick, J. R., & Schupp, J. (2007). *The German Socio-Economic Panel Study (SOEP)—Scope, evolution and enhancements* (SOEP papers on Multidisciplinary Panel Data Research 1). Berlin, Germany: DIW Berlin.

Wooden, M. (2013). *The measurement of cognitive ability in wave 12 of the HILDA survey* (HILDA Project Discussion Paper Series, No. 1/13). Melbourne, Australia: Melbourne Institute of Applied Economic and Social Research.

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*, *22*, 51–68.

Author Biography

Tarek Al Baghal is a research fellow at the Institute of Social and Economic Research, University of Essex, United Kingdom. He is the lead questionnaire designer for the Understanding Society Innovation Panel, involved in designing and implementing this longitudinal study, now on its ninth wave. E-mail: talbag@essex.ac.uk