THE UNIVERSITY OF MELBOURNE
FACULTY OF BUSINESS & ECONOMICS

HILDA PROJECT TECHNICAL PAPER SERIES
No. 1/12, June 2012

The Development of Cognitive Ability Measures in the HILDA Survey

Mark Wooden

with contributions by:

Andrew Mackinnon, University of Melbourne

Bryan Rodgers, Australian National University, and

Tim Windsor, Flinders University

The HILDA Survey Project was initiated, and is funded, by the Australian Government Department of Families, Housing, Community Services and Indigenous Affairs

MELBOURNE INSTITUTE
of Applied Economic and Social Research

ACKNOWLEDGEMENTS

This paper was produced as a by-product of developing the survey instrument for wave 11 of the Household, Income and Labour Dynamics in Australia (HILDA) Survey, a project initiated and funded by the Australian Government Department of Families, Housing, Community Services and Indigenous Affairs (FaHCSIA) and managed by the Melbourne Institute of Applied Economic and Social Research. It uses unit-record data collected during the Dress Rehearsal stage for wave 11. It also has been supported, in small part, by funding from an Australian Research Council (ARC) Discovery Grant (#DP1095497).

The findings and views reported in this paper, however, are those of the authors and should not be attributed to either FaHCSIA, the ARC, or the Melbourne Institute.

The authors are also indebted to Simon Freidin, who produced many of the statistics reported on in this paper, and Nicole Watson, who assisted with some of the analyses.

Introduction

Being a panel study, the majority of the content included in the HILDA Survey is repeated every wave. Provision, however, is also made for rotating content. Within the personal interview component of the survey, and as discussed in Watson and Wooden (2010), the plan going forward is to structure most rotating content around five major topics repeated every four years. These topics cover:

(i) household wealth, previously included in waves 2, 6 and 10;

(ii) family formation and fertility, previously included in waves 5, 8, and 11, and thereafter moving on to a 4-year cycle;

(iii) retirement and retirement expectations, previously included in waves 3, 7 and 11;

(iv) health, included for the first time in wave 9; and

(v) education, skills and abilities, included for the first time in wave 12.

In developing the wave 12 content on education, skills and abilities a major issue has been whether we should include any "tests" of cognitive ability and if so, how much of the content should such tests use.

That cognition is important for behaviour and many socio-economic outcomes is widely accepted. Cognition measures, however, are typically not included in the instruments used in large-scale surveys, which as noted by Lachman and Tun (2008, p. 506), may reflect the common assumption "that reliable and valid assessments are too difficult and time consuming to administer in a survey format by lay interviewers". In some studies, and notably those of youth cohorts (including both the NLSY in the US and LSAY in Australia), these problems can be overcome by linking survey data to results from standardized achievement tests that are administered in the classroom while respondents are still at school. In surveys of wider populations this is usually not possible.

Studies of older cohorts, where the issue of cognitive decline in old age is critically important, however, have long been administering simple tests of cognition using lay interviewers, often over the telephone. A noteworthy example here is the Health and Retirement Study (HRS), which regularly includes measures intended to test the memory, working memory, mental status, reasoning, vocabulary and numeracy of its respondents (see Ofstedal, Fisher & Herzog 2005).

Tests of cognition, however, have also been included in surveys of broader populations. The Brief Test of Adult Cognition by Telephone (BTACT), which involves seven relatively simple to administer tests, was administered as part of the the second wave of the Midlife in the US (MIDUS) study, a telephone survey following up a national sample of Americans aged between 32 and 84 at the time of the second wave (Lachman & Tun 2008).[1] In Australia, the Path Through Life Study, an ongoing, population-based, longitudinal cohort study of young (20-24 years at baseline), midlife (40-44 years at baseline), and older (60-64 years at baseline) adults randomly selected from the Australian Capital Territory and Queanbeyan regions, also regularly includes tests of cognition.

Cognitive tests have also recently been introduced into household panel surveys. The German Socio-Economic Panel included for the first time in its 2006 wave what they described as two ultra-short tests of intellectual ability: (i) a computerized version of the symbol digits

[1] Detailed information about the BTACT can be found at:
http://www.brandeis.edu/departments/psych/lachman/instruments/index.html

modalities test (see Smith 1973); and (ii) a test of verbal (semantic) fluency – the animals naming task (for detailed description of these tests and how they were implemented in the GSOEP, see Lang et al. 2007). Even more recently, a series of cognitive tests were piloted, and eventually included in wave 3 of Understanding Society: The UK Household Longitudinal Survey (see Gray et al. 2011).

Issues

There are a number of obstacles to, and problems with, the inclusion in the HILDA Survey of measures designed to assess the level of cognitive ability among respondents.

(i) The measures could easily use up all (and more) of the interview time set aside for rotating content (typically 10 minutes). The BTACT, for example, despite being described as a short test, still takes "less than [that is, up to] 20 minutes" (Lachman & Tun, p. 515).

(ii) Many of the available tests are complicated to administer, and certainly not easily administered over the telephone (bearing in mind that between 8% and 9% of HILDA Survey respondents are interviewed by phone). That said, tests suitable for administration by the telephone are now well established, as reflected in both the BTACT and the cognitive function component of the HRS. And with a sample as large as the HILDA Survey it does not necessarily follow that all respondents have to be administered all parts of the survey.

(iii) Many respondents may be averse to questions that come across as "tests", which may, in turn, have ramifications for attrition at later waves.

(iv) Performance can be affected by motivation (especially for those tests where there is a time limit), but we are loathe to motivate performance by placing pressures on participants to perform.

(v) Given tasks will all be administered in English, scores will reflect both cognitive ability and English language ability. This will be especially so for any tasks measuring some aspect of verbal intelligence or capacity.

(vi) Performance on some tasks can be influenced by observing other household members completing the task. This suggests the need to either ensure tasks are conducted in private (which is not possible in some households), administer tasks in a way that they are not easily observed by others (e.g., self-administered), or have multiple forms of the same task.

(vii) The administration of some tests may be hindered by copyright and other restrictions, and / or require the payment of substantial license fees.

Short-listing Possible Measures

Bryan Rodgers (ANU), a member of the HILDA Survey External Reference Group, in collaboration with Tim Windsor (a psychologist at Flinders University), prepared a short paper summarising four measures they recommended be piloted in the HILDA Survey. This is provided as Appendix A.

In addition, consideration was given to the range of cognitive ability measures that were being included in other longitudinal studies. A summary listing of the various tests of cognitive function that were under consideration is provided in Table 1. This table also indicates which tests have been advocated by Rodgers and Windsor, and which tests have been used in the BTACT, HRS, Understanding Society (USoc) and GSOEP. A short simple description of each measure is provided in Table 2

Table 1: Alternative Cognitive Ability Measures under Consideration

Ability	*Test*	*Recom by:*	*Used in:*			
		Rodgers & Windsor	*BTACT*	*HRS(a)*	*USoc*	*GSOEP*
Episodic verbal memory	Word list recall – immediate		X	X	X	
	Word list recall – delayed		X	X	X	
Working memory span	Backwards digit span	X	X			
	Series 7s subtraction			X	X	
Verbal fluency	Category fluency	X	X	X	X	X
Attention switching / reaction time	Red / green test		X			
Inductive reasoning	Number series		X	X	X	
Processing speed	Backwards counting		X	X		
	Symbol digits modalities	X				X
Verbal intelligence	Spot the word	X				
	Vocabulary (adapted from WAIS-R)			X		
	National Adult Reading Test (NART)					
Numeric ability	Simple "everyday" arithmetic problems			X	X	

Note: (a) The HRS also asks older respondents to name the date, to name simple objects, and to name the current President and Vice-president of the USA.

Rodgers and Windsor argued that we needed measures of both fluid and crystallized ability (or intelligence) and that for the piloting stages we should include at least two tests of each type. Taking into account both ease of administration and administration time and prevalence of use in the fields of development psychology and cognitive ageing, they proposed the inclusion of four measures: Backwards Digits Span; Category Fluency; Symbol Digits Modalities; and Spot the Word. To that list we then added the number series task (as administered in the BTACT).

It was subsequently decided that we should not proceed with Spot the Word. This reflected concerns about the future and uncertain format of this test (the owners of the test indicated that the test was undergoing a major redevelopment), its cost to administer (in terms of license fees), and copyright restrictions that might preclude adaptation of the test. In its place we included the National Adult Reading Test (NART), even though this particular test was not on our original list.

Table 2: Summary Description of Alternative Cognitive Ability Measures

Test	*Brief description*	*Time (secs)*	*Mode*	*Notes*
Word list recall – immediate	Series of words (10 or 15) is read out and respondent to ask to repeat as many as they can recall.	90	Oral	BTACT uses a list of 15 words. USoc uses a list of 10 words.
Word list recall – delayed	Respondent asked to recall words from same list in test above, but after a period of 10 to 15 minutes has elapsed.	40	Oral	The number of correct items, repetitions and intrusions are recorded.
Backwards digit span	Respondent hears increasingly longer series of digits (ranging from 2 to 8 digits) and asked to repeat them in reverse order.	90 (ave)	Oral	Test is discontinued if both trials at a set size are missed.
Series 7s subtraction	Respondent asked to begin at 100 and subtract 7. The test involves five serial subtractions.	40	Oral	Scoring is not so obvious.
Category fluency	Respondent asked to name as many animals as they can in one minute.	100	Oral	Measurement error likely to be pronounced in absence of interview recording.
Red / green test	Respondents are asked to respond “stop” and “go” when they hear the words “green” and “red”. The procedure is then repeated with response options reversed. The procedure is repeated a third time with respondents given cues to switch back and forth between these two modes.	200	Self-admin or oral	Usually this test would be administered by CAPI. Over the telephone is much more difficult and would almost certainly require interviews to be recorded.
Number series	Respondents are asked to complete a number sequence (BTACT) or fill in a missing number in a sequence (USoc)	150 to 250	Oral or written	The USoc version requires respondents to write the number sequence down.
Backwards counting	Respondents are asked to count backwards from 100 as quickly as possible in 30 seconds.	45	Oral	Score is total number of correct numbers reported.
Symbol digits modalities	Respondents use a key to pair up as many symbols with digits as they can in 90 seconds.	180	Oral or written	The GSOEP version of this test is designed for delivery by computer.
Spot the word	Real words and nonsense words are paired and respondent has to choose the real word.	240	Written	Standard test involves 60 pairs.
Vocabulary	Respondents are asked to define he meaning of series of words	100	Oral	Test is adapted from WAIS-R, which involves a list of 35 words. HRS uses just five words.
NART	Respondents are asked to read out a list of 50 irregularly spelt words	180	Oral	Can only be administered in-person.

Test	*Brief description*	*Time (secs)*	*Mode*	*Notes*
Simple "everyday" arithmetic problems	Up to six mathematical problems set in an everyday context are asked.	100 to 300	Oral	All respondents get asked three questions, the answers to which determine subsequent routing.

The Tests

Backwards Digit Span (BDS)

This is a test of working memory span, which, as noted by Rodgers and Windsor, features in many traditional intelligence tests, and most notably the Wechsler Adult Intelligence Scales. The format proposed for the HILDA Survey is taken directly from the BTACT but is equally suitable for delivery in face-to-face interviews. It involves interviewers slowly reading out successively longer strings of single-digit numbers and asking participants to repeat those strings in reverse order. Respondents are given two chances at each length or 'level'. When the respondent gets one trial correct at a level, the first trial at the next level is administered. If the first trial is incorrect, the second trial is administered. If both responses at the same level are incorrect, the test is discontinued. The longest sequence administered is eight digits.

The main potential problems with the administration of the BDS task in the HILDA Survey is the household context and hence the likelihood of both distractions adversely affecting outcomes (given number sequences cannot be repeated) and learning effects from one household member seeing another household member undertake the task. There is also the possibility that telephone respondents could use pen and paper to assist them (despite instructions not to do so).[2]

Category Fluency

This is a test of verbal fluency which seeks to "evaluate the spontaneous production of words under restricted search conditions" (Strauss, Sherman & Spreen 2006, p. 499). There are numerous forms of this test. The form recommended for the HILDA Survey is the one that seems most common in general social science surveys, and requires respondents naming as many things from a specified category as possible within a one-minute interval. The most common category is 'animals', though other categories such as 'foods' and 'clothing' are also used.

For the HILDA Survey we also use animals. Names of extinct, imaginary or magic animals are admissible, but given names like 'Fido' are not.

A potentially major weakness of this test is that interviewers are required to record all answers so that they can then identify (and exclude) repetitions. There is thus the possibility that scores may be a function of the speed of interviewers recording answers.

There is also again the potential for intra-household earning effects; indeed, it is much more likely with this test. For that reason we have altered the category for sample members who have previously witnessed the test being conducted to 'food'.

[2] The experience with the implementation of the BTACT (see Lachman & Tun 2008) suggests no marked difference differences in the performance of telephone and face-to-face respondents on any of the tests administered.

Number Series

This is a test of inductive reasoning, but which requires a base level of numeracy. Respondents are read out a series of five numbers which have a logical pattern and then asked to nominate the next number in the sequence. The version we use is taken directly from the BTACT, which in turn comes from Salthouse and Prill (1987), and involves a total of five items.

As with the previous tests there is the potential for learning by other household members.

Symbol Digits Modalities (SDM)

This test was originally developed as a screening measure for cerebral dysfunction, but has been widely used in broader settings as a general test for divided attention, visual scanning and motor speed (Strauss et al. 2006, p. 617).

The test involves participants matching symbols to numbers using a printed key. The score is the number of items correctly matched within a 90 second time interval.

The test can be administered either orally or in written form. For the HILDA Survey we are adopting the latter. It is not suitable for telephone delivery and so will not be administered to respondents interviewed by phone. Further details about the test are provided in the test manual (Smith 2007).

The test is protected by copyright, which is owned by Western Psychological Services (WPS) in the US. Permission to use the test in wave 12 of the HILDA Survey has been obtained from WPS.

In terms of administration the main issue is the lengthy and complicated instructions that the interviewer is required to read out prior to the test commencing.

National Adult Reading Test (NART)

The NART is a reading test of 50 irregularly spelled words, listed roughly in order of difficulty, which is intended to provide an estimate of pre-morbid intelligence. The value of the test lies, in part, in the high correlation between reading ability and intelligence in the normal population, with numerous studies reporting moderate to high correlations between NART performance and measures of intellectual status (see Strauss et al. 2006). Indeed, scores on the NART are designed to predict scores on the WAIS-R intelligence test. Further details about the test rationale, as well as procedures for administration and scoring are provided in the test manual (Nelson 1982).

The test is designed to be administered to persons aged 18 years or over, though in Australia seems to have been mainly used in studies of older populations (e.g., Kiely et al. 2011). In the HILDA Survey we will be administering it to all persons in our sample capable of reading English, which will include persons as young as 15 years.

An obvious weakness with the test is that it is only intended for persons who can read English. It is thus not an appropriate measure of intelligence for non-English language speakers or for persons whose reading ability has been seriously compromised by injury or illness. Nevertheless, given we are also interested in reading ability in its own right, there are still good reasons to administer the test to non-native English language speakers in the HILDA Survey sample. That is, we anticipate that NART scores could also double as a measure of functional literacy.

The test involves participants being presented with a word card and instructed to read out loud each word. Interviewers then record correct pronunciations, with the total correct

providing the score. Slight variations in pronunciations due to regional accents are acceptable. Again it is not easily delivered over the telephone, and hence will only be administered to HILDA Survey respondents interviewed face-to-face.

A major issue is to determine which pronunciations are permissible, with many dictionaries allowing different versions of some words. For the HILDA Survey we will be permitting pronunciations that are regularly used in not just Australia (as reflected in the Macquarie Dictionary), but also in the UK and the US.

Administration of the test clearly places considerable burden on interviewers to be able to identify correct and incorrect pronunciations. This will be facilitated by training and the provision of pronunciation guides (both in written and audio formats). Previous research indicates that inter-rater reliability is actually very high – typically above 0.88 (Strauss et al. 2006, p. 196). However, it is possible, indeed likely, that the test administrators used in these other studies are very different to the average interviewer employed on the HILDA Survey.

Pilot Testing

All five measures, along with the introductory script and additional questions collecting para-data, were 'skirmish tested' in October 2011 on a group of 31 people by staff from the fieldwork provider, Roy Morgan Research. The skirmish test sought to identify both obvious problems with the question script and sequence and participant reactions to the measures.

The script was then amended and embedded within the larger computer-assisted survey instrument in readiness for the annual dress rehearsal.

Each wave of the HILDA Survey is preceded by a Dress Rehearsal (DR), conducted each year in March and April, on a separate sample that is also being followed over time. As the name implies, the intent of the Dress Rehearsal is to trial all procedures and instruments to be implemented in the main survey. There are, however, important differences. First, and most notably the fieldwork window for the DR is relatively short – just two months compared with over 6 months in the main survey. Second, the sample for the DR is both relatively small and geographically concentrated – the original selections were restricted to households in New South Wales and Victoria, and mainly in Sydney and Melbourne. Third, the survey instruments used in the DR are typically much longer than used in the main survey, and the wave 12 DR is no exception.

For wave 12, 834 households from this panel sample were issued to field, with 697 of those households participating during the two-month fieldwork window. The total number of completed individual interviews generated was 1334.

A paper version of the script used to deliver the cognitive ability tests in the DR is provided in Appendix B. As can be seen, there is an introductory script explaining the purpose of the tests, or what is referred to as "exercises". Each of the individual tests is preceded by a question clarifying that the respondent understands the task and providing an opportunity for respondents to decline to undertake that task.

There are also questions asking interviewers to record: whether each test was commenced, and if not why not; whether the test was completed, and if not why not; and the presence of other persons during the test, and whether they assisted;

Dress Rehearsal Outcomes

Willingness (and ability) to participate

Given its expected length, participants were reminded at the start of this section that answering these questions, like all questions in the HILDA Survey, is entirely voluntary. However, we also included in the script an instruction requesting interviewers record whether it was okay to proceed. It seems that many interviewers treated this as a prompt for opting out, and indeed often specifically checked with participants whether they were okay to continue. As a result, just over 16% of participants (n=215) were recorded as requesting not to do the tasks, an unexpectedly high figure.

Further, this group is a self-selected group that differs from the "willing" participants in some very obvious ways. This can be seen in Table 3, which reports the proportion of persons indicating they wished to skip these tasks cross-classified by selected characteristics. As can be seen, persons that chose not to participate in the cognitive tasks were more likely to: be elderly (over 65 years); have not completed 12 years of schooling (which will, in turn, be a function of age); have a serious work-limiting long-term health condition (though such persons represent a fairly small proportion of the sample - about 2.8%); and report having relatively poor English language and numeracy skills.

Such obvious sources of selectivity bias suggest that the script should be revised to remove the scope for interviewers to encourage participants to opt out of this sequence. On the other hand, there is the potential to create unwanted stress and anxiety for some respondents.

Given the reduction in the number of tasks to be administered in the main survey together with interviewer feedback that those respondents that did participate typically found the tasks a refreshing change (and especially the Category Fluency, NART and Symbol Digits Modalities tasks), our view was to opt for a strategy that encourages greater participation in these tasks.

For the main survey, therefore, we reworded the introductory text to remove the reminder about the voluntary nature of participation. Thus the introduction to this section in the Main Survey will read:

> "A special feature of the interview this year is the next section. It comprises three short exercises that involve you remembering and making judgments about words, symbols and numbers."

Participants also had the opportunity to opt out prior to each task. More specifically, each task began with a set of instructions that was read out to each participant, which effectively provided the respondent a further chance to opt out. In addition, respondents may fail to complete tasks, though in most cases their scores obtained at the time of discontinuation would still be treated as valid.

Table 4 provides some simple descriptive data summarising the incidence of these non-starts and discontinuation by task. The Category Fluency task seems to have been the best received while Number Series was the worst received. This is in line with the subjective assessments of interviewers. Note that comparisons across tasks are complicated by order effects. All tasks were completed in the order shown in Table 4 and it seems likely that interest in completing the tasks may have waned over the course of the testing (which, as discussed later, averaged over 17 minutes for respondents completing all five tasks). As a result, the relatively high proportion of non-starters for the NART and SDM tasks may be misleading.

Table 3: Percentage of sample that chose not to undertake cognitive tasks, by selected characteristics

Characteristic	*%*	*Chi-squared*	*P diff = 0*
Sex		0.14	ns
Men	15.7		
Women	16.5		
Age group		25.24	<.001
<25	10.5		
25-34	13.9		
35-44	17.3		
45-54	13.7		
55-64	16.4		
65-74	27.4		
75+	25.5		
Education level		11.77	.008
Less than Year 12	21.0		
Year 12	18.9		
Trade qualification / Diploma	14.9		
Degree or higher	12.0		
Long-term health condition		13.57	.004
None	15.5		
Mild (does not limit work)	17.1		
Moderate	14.6		
Severe (cannot work at all)	37.8		
Family status		10.95	.012
Partnered w children	17.9		
Partnered, no children	20.4		
Single w children	12.2		
Single, no children	13.2		
Employment + hours of work per week		8.03	.045
Not employed	19.3		
<35 hours	12.1		
35-49 hours	15.9		
50 hours or more	13.9		
English speaking ability		89.38	<.001
Only speaks English at home	13.6		
Very well	16.3		
Well	35.4		
Not well / Not at all	76.0		
English reading ability		86.39	<.001
Excellent	11.8		
Good	16.4		
Moderate	33.0		
Poor	61.1		
Mathematical skills		29.41	<.001
Excellent	11.9		
Good	14.2		
Moderate	20.2		
Poor	34.7		

Table 4: Participation outcomes, cognitive ability tasks (Ns, unless otherwise stated)

	Category fluency	*BDS*	*Number series*	*NART*	*SDM*
Started task	1102	1090	1052	944	939
Did not start task	17	29	67	44	49
(% of "willing" participants)	(1.5)	(2.6)	(6.0)	(4.5)	(5.0)
Task not completed in full	4	24	55	18	5
(% of all participants starting)	(0.4)	(2.2)	(5.2)	(1.9)	(0.5)

Time taken

The section on cognitive ability averaged 13.3 minutes across all respondents. For persons that completed all tasks, it was much longer – almost 18 minutes. The overall length of the main individual survey instrument (the Continuing Person questionnaire or CPQ) administered in the DR was 44.3 minutes, which is well in excess of our contractual target – just 35 minutes. Thus as anticipated we will be unable to include all five cognitive ability measures in the main survey, at least not without making significant changes to other ongoing (annual) content or exceeding our target.

The breakdown of time taken by task is provided in Table 5. To keep the CPQ within (or close to) the target length in the main survey we will be unable to devote more than 8 minutes of space in the main survey instrument to cognitive ability tasks (depending on decisions made about other survey content).

Table 5: Average interview times (minutes), cognitive ability tasks

Sub-section	*Persons that completed the sub-section*	*All persons*
Introduction	0.8	0.8
Category Fluency	2.7	2.2
Backwards Digit Span	2.2	1.8
Number Series	3.5	2.4
NART	3.8	2.7
Symbol Digits Modalities	4.8	3.4
Total Length	17.8	13.3

Given a marked reduction in the length of time taken to introduce the tasks, which we think will now average no more than 12 seconds (i.e., 0.2 of a minute), and taking into account an assumed increase in the proportion of participants that are administered the tasks (to 90% for tasks that can be done on the telephone and 80% for those that cannot), our assessment is that the following combinations would be associated with the following expected interview times:

- Category Fluency + BDS = 4.6 minutes
- BDS + NART = 5.2
- BDS + SDM = 6.0

- Category Fluency + SDM = 6.4
- NART + SDM = 7.0
- Category Fluency + BDS + SDM = 8.4
- BDS + NART + SDM = 9.0

As explained later, however, we could include a markedly shorter version of the NART, which would reduce all times associated with the NART by an estimated 1.4 minutes.

Descriptive statistics

Summary descriptive statistics for the test scores on each of the five cognitive ability tasks are presented in Table 6. The test scores for each are derived as follows:

- Category Fluency is scored as the total number of words recorded during a 60 second period less any repeated words and less any "intrusions" (words that do not fall within the stated category, which for most participants is "animals").
- The score for the Backwards Digit Span is the highest number of digits correctly recalled and repeated in reverse order. It will usually range between 2 and 8 (though zero is also a valid score).
- The Number Series task involves correctly identifying the next number in a five number arithmetic sequence. There are a total of five sequences and hence scores can range from 0 to 5.
- The National Adult Reading Test (NART) requires correctly pronouncing 50 irregularly spelt words from a card and hence scores can range from 0 to 50.
- The Symbol Digits Modalities (SDM) task involves matching numbers to symbols using a keycard. The score is simply the number of correct matches achieved within a 90 second time frame.

As can be seen, with the exception of Number Series, all of the tests have properties that suggest distributions that are close to normal. Category Fluency and Backwards Digit Span have distributions that are somewhat skewed to the left, while scores on both NART and SDM are slightly skewed to the right.

It is also clear that while all test scores are positively correlated with each other, the size of that correlation is not so large to suggest that any of the tasks are redundant. This can be seen in Table 7. The lowest correlation – .18 – is between SDM and NART, suggesting these are the least substitutable. Once we adjust for age, however, this correlation almost doubles in size, with the smallest correlation now between Category Fluency and Backwards Digit Span.

We next checked whether cognitive ability scores vary in expected ways with both age and education level. Thus in Table 8 we present data on mean scores cross-classified by both age group and a crude (binary) indicator of educational attainment. As expected we can see that persons with at least some post-school education score, on average, better than persons with lesser education levels. That said, these differences are only pronounced for persons aged 25 to 74 years, and even then may not be as large as might have been expected.

Table 6: Summary descriptive statistics, cognitive ability tasks

Statistic	*Category fluency*	*BDS*	*Number series*	*NART*	*SDM*
Mean	20.99	4.97	2.80	29.25	49.76
Median	20.00	5.00	3.00	30.00	50.00
Std. deviation	6.34	1.55	1.63	9.93	13.37
Minimum	4	2	0	1	4
Maximum	47	8	5	49	95
Skewness	.48	.41	-.29	-.44	-.21
Kurtosis	.46	-.69	-1.11	-.19	.20
N	1102	1090	1052	944	934

Table 7: Correlation between cognitive ability scores (age-adjusted correlations in parentheses)

	Category fluency	*BDS*	*Number series*	*NART*	*SDM*
Category Fluency	1	.24 (.20)	.32 (.31)	.29 (.35)	.41 (.36)
Backwards Digit Span		1	.43 (.42)	.36 (.38)	.36 (.34)
Number Series			1	.40 (.45)	.40 (.38)
NART				1	.18 (.35)
SDM					1

We can also see that scores tend to decline with age, though typically these differences only become pronounced in old age (after 65). The SDM scores exhibit a slightly different pattern, peaking at quite a young age (in the 25 to 34 year range) and then dropping quite rapidly during prime-age as well as older age. The biggest outlier, however, is NART, where test scores tend to rise with age before falling in the 65+ year age range. We suggest this reflects NART's superiority as a measure of crystallized intelligence, which tends to improve with experience, whereas most of the other tests are measures of fluid intelligence. It also suggests that the Category Fluency task may not be the good measure of crystallized intelligence that Rodgers and Windsor suggest. Most obviously, the fact that the test imposes a time constraint means that working memory comes into play, and working memory is much more closely related to fluid intelligence.

In Table 9 we present evidence on the relationship between cognitive ability scores and self-assessed indicators of proficiency in English literacy and numeracy. In almost all cases, mean test scores are markedly lower for those that self-assess as having relatively poor skills. The one exception is Category Fluency and mathematical skills, where there is no relationship, and which is entirely consistent with expectations. Such results are reassuring, though arguably tell us more about the value of self-assessed data than they do the value of the cognitive ability tests.

Table 8: Mean cognitive ability scores by age group and education level

Age group	*Education level*	*Category fluency*	*BDS*	*Number series*	*NART*	*SDM*
<18	High school	20.9	4.9	2.7	23.0	53.8
	Post-school	*	*	*	*	*
	Total	21.1	4.9	2.8	22.8	53.6
18-24	High school	22.1	5.3	3.3	27.0	58.9
	Post-school	21.9	5.3	2.6	25.5	54.8
	Total	22.0	5.3	2.9	26.1	56.6
25-34	High school	20.1	4.7	2.7	23.7	54.7
	Post-school	22.8	5.3	3.3	29.5	57.5
	Total	22.4	5.2	3.2	28.8	57.2
35-44	High school	20.4	4.9	3.0	26.9	52.6
	Post-school	22.7	5.0	2.9	29.5	54.1
	Total	22.1	5.0	2.9	28.8	53.7
45-54	High school	21.2	4.6	2.5	27.4	45.2
	Post-school	21.7	5.2	3.1	32.2	51.2
	Total	21.5	5.0	2.9	30.8	49.5
55-64	High school	18.5	4.3	2.2	28.4	38.9
	Post-school	20.8	5.1	3.0	35.1	46.5
	Total	20.1	4.8	2.7	33.0	44.3
65-74	High school	17.2	4.1	1.9	26.5	36.6
	Post-school	20.4	4.9	2.6	36.1	39.7
	Total	18.8	4.5	2.3	31.3	38.1
75+	High school	15.0	4.2	1.3	30.6	31.4
	Post-school	16.6	4.8	1.9	32.3	33.4
	Total	15.6	4.4	1.5	31.3	32.2
All persons	High school	19.7	4.7	2.5	26.5	47.4
	Post-school	21.8	5.1	3.0	31.0	51.2
	Total	21.0	5.0	2.8	29.3	49.8

We would also expect that scores on the NART would be much more strongly associated with English language ability than the other tests. In fact, it is the test that ex ante we would have thought was least affected – Number Series – where the scores were most strongly correlated. If we look at English speaking ability and compare non-native speakers who report they speak English well with those that report they speak English very well, the ratio in mean scores is 0.80 for Category Fluency, 0.83 for BDS, 0.78 for SDM and NART, and just 0.54 for Number Series.

When you compare persons with moderate English reading skills with those with excellent skills the ratios are 0.79 for Category Fluency, 0.79 for BDS, 0.88 for SDM, 0.55 for NART and 0.61 for Number Series. As we would expect, the inverse relationship between English language skills and NART is much stronger in respect of reading ability.

Finally, we examined associations with the hourly wage (for employees). Correlation coefficients are reported in Table 12 and show, after removing outliers (cases where the derived hourly wage was less than $10), reasonably high positive correlations with the hourly wage for all tasks. This is especially so after controlling for age, sex and education.

Table 9: Mean cognitive ability scores by self-assessed measures of English language ability and numeracy

	Category fluency	*BDS*	*Number series*	*NART*	*SDM*
English speaking ability					
Only speaks English at home	21.4	5.0	2.9	29.9	49.9
Very well	19.7	5.2	2.8	27.8	53.3
Well	15.7	4.3	1.5	21.8	41.7
Not well / Not at all	*	*	*	*	*
English reading ability					
Excellent	22.2	5.2	3.1	32.7	52.7
Good	19.3	4.6	2.3	24.5	44.9
Moderate	17.6	4.1	1.9	18.1	46.4
Poor	16.6	3.5	0.5	*	*
Mathematical skills					
Excellent	22.7	5.4	3.5	32.9	54.8
Good	20.5	4.9	2.8	28.8	48.8
Moderate	19.5	4.6	2.0	26.9	46.4
Poor	21.76	4.2	1.3	20.6	41.7

Table 10: Correlations between hourly wage and cognitive ability scores among employees

Cognitive ability task	*r*	*Partial r**
Category Fluency	.177	.236
Backwards Digit Span	.067	.132
Number Series	.102	.170
NART	.258	.210
Symbol Digits Modalities	.111	.235

* Controlling for sex, age and education.

Comparisons

It would be informative to know whether the data collected are consistent with data collected in other studies, and for some tests (but not all) there are readily available comparisons, though all are far from ideal.

Category Fluency. While we are not aware of any large scale Australian study that has implemented this task, there are data available for other countries. In Table 11, for example, we report comparisons with a small Canadian sample, and perhaps surprisingly, given differences in sample size, timing and institutions, find remarkably similar distributions in scores.

Table 11: Category Fluency – Comparison with normative data for Canada (persons aged 16 to 59 years)

	HILDA W12 DR		*Canadian study (Tombaugh, Kozak & Rees 1999)*	
	9-12 years of education	*13+ years of education*	*9-12 years of education*	*13+ years of education*
Percentile score				
90	29	30	26	30
75	24	26	23	25
50	20	21	20	23
25	16	17	17	18
10	13	14	15	16
Mean	20.2	21.8	19.8	21.9
(SD)	(5.9)	(6.4)	(4.2)	(5.4)
N	369	684	109	78

A German language version of this test was also implemented in the 2006 wave of the German Socio-Economic Panel (Anger and Heineck 2008). The mean scores reported in the GSOEP data are noticeably higher than recorded in the HILDA Survey Dress Rehearsal – 26 vs 21 – and there is also a longer upper tail in the former – maximum score of 74 compared with 47 in our test. The most likely explanation for these differences is the obvious differences in the way the tests were administered (rather than the language difference). The GSOEP adaptation of this test involved interviewers simply clicking on a button on a computer keyboard (see Lang et al. 2007), whereas we followed the traditional version which requires answers to be written down by the interviewer. The German approach can obviously lead to errors given interviewers have to instantly identify repeated words. The traditional approach, however, has the obvious (and we would argue, more serious) disadvantage that scores can be correlated with the speed of the interviewer.

BDS. This test has been implemented during the Path Through Life study. Summary results from the first wave, disaggregated by gender, are reported in Jorm et al. (2004), and replicated below in Table 12. Also provided are comparable data from the HILDA DR.

As can be seen, scores appear to be noticeably lower in the HILDA sample for men in all age cohorts, and among women in the 40-44 year groups, though it is only among the 40 to 44 year olds that these differences are statistically significant. Higher scores might be expected in the PATH sample given its restriction to residents of ACT and Queanbeyan, and hence more highly educated sample members. Less obvious is why the differences should only be marked for one of the three age cohorts.

SDM. Summary data from a relatively small US sample (from the mid-1970s) are reported in the test manual (Smith 2007) and replicated below in Table 13, along with comparable data from the wave 12 DR. While the test manual refers to these data as providing 'population norms', in no sense could the sample used be seen as representative of the US population. Indeed, if the SDM is administered in our main sample, this may well be the first time anywhere that this test has been administered to a nationally representative population. Note

also that the US data were collected in clinical settings which will almost certainly mean differences with a sample where subjects are tested in their homes.

Despite these obvious differences, the HILDA DR data provides mean scores that follow broadly similar patterns, especially within the age range (25 to 64), and especially among the more highly educated.

Table 12: Mean scores (standard deviations in parentheses) on BDS – comparison with Path Through Life

Age group	*HILDA W12 DR*		*PATH*	
	Men	*Women*	*Men*	*Women*
20-24	5.22 (1.70)	5.17 (1.54)	5.47 (2.31)	5.23 (2.26)
40-44	4.84 (1.46)	4.57 (1.28)	5.35 (2.36)	5.10 (2.23)
60-64	4.80 (1.21)	4.78 (1.67)	5.00 (2.26)	4.75 (2.22)

Table 13: Mean scores (standard deviations in parentheses) on written SDMT – comparison with normative data for USA

Age group	*HILDA W12 DR*		*US (Smith 2007)*	
	12 or less	*13 or more*	*12 or less*	*13 or more*
18-24	58.94 (9.79)	54.77 (11.17)	54.40 (8.31)	61.93 (10.15)
25-34	54.72 (10.95)	57.48 (11.46)	53.30 (7.98)	57.72 (9.08)
35-44	52.58 (11.74)	54.09 (12.06)	51.50 (8.03)	54.20 (11.17)
45-54	45.20 (10.26)	51.16 (9.69)	47.26 (9.56)	52.27 (8.48)
55-64	38.90 (11.18)	46.50 (9.51)	42.80 (8.08)	47.60 (8.31)
65-78	37.37 (12.85)	38.35 (10.34)	33.31 (9.02)	43.55 (11.27)

The SDM test has also been administered in the PATH Through Life Project (Jorm et al. 2004). In that study the scores are noticeably higher than we find (by 6 to 8 points). This almost certainly reflects differences in administration – in PATH the SDM test was administered in its oral form rather than the written form used in the HILDA Survey DR, and the test score norms for oral administration are 6 to 8 points higher than for written administration.

NART. While the NART has been extensively used, data from large population-wide samples have proven difficult to locate. In Australia, for example, Kiely et al. (2011) report NART norms from three separate studies, all of which cover elderly populations (65 years or older).

Influence of External Circumstances

Ideally all of these tests should be administered in a quiet environment with no one else present, but in surveys conducted in private homes this is often not possible. As shown in Table 14, somewhere between 16% and 18% of interviews were conducted with someone else present at the time the cognitive ability tasks were conducted.

In all cases, except the SDM, the presence of another adult was associated with a lower test score, suggesting the presence of others was a distraction (though only in the case of the NART was the difference statistically significant at conventional levels; t=3.88).[3] In the case of the SDM, persons with someone else present appeared to perform better (mean = 52.2 vs 49.7) but the difference was a long way from achieving statistical significance.

More problematic is the possibility that another household member yet to be interviewed witnesses the test and as a result is better prepared and so performs better. This can affect all test scores but is arguably most problematic for the Category Fluency and Number Series tasks.

We attempted to deal with this problem in the case of Category Fluency by assigning respondents who have previously witnessed the test a different category – foods, rather than animals. This of course means that scores for respondents naming foods are not strictly comparable with participants naming animals. The proportion administered the "foods" category was relatively small – just 5% of all persons taking this test. However, the mean scores for this group were significantly higher – 23.5 vs 20.9 (t=2.85). Whether this outcome is because it is easier to name foods or because of learning effects from seeing the test administered previously (albeit with a different category) is unclear.

In the case of Number Series no actions were taken to prevent or inhibit learning effects.

Table 14: Presence of others during cognitive ability tasks (%)

	Category fluency	*BDS*	*Number series*	*NART*	*SDM*
No one present	82.8	83.9	83.7	82.4	83.5
Another sample member	15.2	13.9	13.8	15.6	14.6
A child under 15	1.8	1.9	2.0	1.6	1.8
A non-household member	1.0	1.0	1.0	1.0	0.7

[3] Weakly significant results (at the 90% confidence level) were found for both Backwards Digit Span and Number Series.

Mode Effects

In the case of three tasks – Category Fluency, BDS and Number Series – administration by both face-to-face methods and telephone was permitted, raising the possibility that scores might be sensitive to mode. There is, for example, greater scope for participants to write numbers down which would lead to superior scores on both the BDS and Number Series tasks.

As can be seen from Table 16, the telephone respondents do score better on average, but the differences are mostly small.[4] Nevertheless, the difference on the Number Series task is large enough to be statistically significant at the 95% confidence level.

Table 15: Mean cognitive ability scores (and standard deviations in parentheses) by survey mode

Survey Mode	*Category fluency*	*BDS*	*Number series*
Telephone	21.57 (7.01)	5.09 (1.54)	3.27 (1.57)
Face-to-face	20.93 (6.23)	4.96 (1.55)	2.75 (1.62)
T-test for significance of difference in means	0.99	0.90	3.46

Interviewer Effects

An issue of some concern, especially for the NART, but also for Category Fluency, is the possibility that test scores are correlated with interviewer ability. A test of interviewers conducted during Dress Rehearsal training revealed relatively poor performance, with an average error of two items, but a number of notable outliers (the maximum number of errors was 14) all of whom were not native English speakers.

Identifying interviewer effects, however, is complicated by the fact that respondents are not randomly distributed across interviewers. Interviewers are allocated workloads that are geographically clustered and hence mean ability scores when averaged across interviewers will be affected by any factor that is both correlated with location and ability (e.g., socio-economic status). Thus we do expect systematic variation by interviewer, but because of factors that are correlated with location, and not because of systematic biases caused by the interviewer.

However, we can take advantage of the strong likelihood that the scope for interviewer effects is considerably reduced in some of the tasks. Most obviously, the Symbol Digits Modalities task is largely completed by the respondent with minimal interviewer involvement. The interviewer's only responsibilities are to read out the instructions and to monitor compliance with the 90 seconds time limit. We would expect very little of the

[4] Given concerns that telephone respondents may have been writing down numbers during the BDS, we also checked for differences in the proportion of extreme values recorded (i.e., scores of 8) and again differences (10.9% vs 8.1%) were statistically insignificant.

observed variation between interviewers in SDM scores to be due to variance in interviewer ability or testing methods.

We thus report, in Table 16, the results from a simple one-way analysis of the variance in cognitive ability scores where the independent variable is interviewer identity. We restrict our data to test scores collected by face-to-face interviewers and only include interviewers that conducted a minimum of 10 interviews (n=23). As can be seen, for most tasks relatively little of the variance is due to differences in scores across interviewers. It is just 6% for the SDM task, rising to 8% for Category Fluency and Number Series. As expected, however, NART is an outlier, with 18% of the variance due to between interviewer effects. This finding is suggestive of relatively poor inter-interviewer reliability in scoring. That said, this is only a weak conclusion. It is still possible that NART scores are much more affected by factors associated with location (such as socio-economic status) than are the other cognitive ability tasks.

Table 16: Analysis of variance in cognitive ability scores by interviewer

% of sum of squared residuals due to:	*Category fluency*	*BDS*	*Number series*	*NART*	*SDM*
Within interviewer effects	92.1	93.7	92.0	82.1	94.0
Between interviewer effects	7.9	6.3	7.8	17.9	6.0
F	3.61	2.80	3.39	8.86	2.59

A Shorter Version of NART: Proposal I

Beardsall and Brayne (1990) have proposed a short version of NART wherein a person is only given the full NART if they get at least 21 out of the first 25 words correct, and their full NART score is predicted from their answers to the first half. The main motivation here is to reduce anxiety or distress for people with poor reading skills.

Based on our DR sample this would have the effect of halving the test length for only 20% of the sample. Further, is the degree of error in predicted values acceptable? A simple exercise comparing actual full NART scores with the predicted scores using the formula proposed by Beardsall and Brayne gives a mean error of 3.4. Moreover, the number of errors is only noticeably lower at the very low end of the distribution (scores of less than 5) where there are very few cases.

There may be good reasons to suspend the NART for persons scoring badly (e.g., who score say 10 or less in the first half of NART), but this will come at the cost of a considerable loss of precision in scores. Some persons who score badly in the first half of NART still manage to correctly pronounce words in the second half of the test.

A Shorter Version of NART: Proposal II

We also considered whether a reliable and valid measure could be derived using only a subset of the words included in the NART. For this exercise we engaged the services of Andrew Mackinnon (from the Orygen Research Centre, University of Melbourne), who used item response theory models to obtain information about item fit and test performance with a view to identifying candidates for removal. A preferred list of 25 items was targeted.

The final list of items for a 25-item version of NART based on the DR data is provided in Table 17. The items are listed in descending order of difficulty, where difficulty is based on the percentage of correct pronunciations recorded in the W12 DR.

This short-form scale proves to be highly reliable (Cronbach's alpha=.89) and scores are highly correlated with scores on the full list (r=.97).

In addition to being much shorter, and thus less burdensome for respondents and interviewers alike, it also has the added benefit that fewer of the items have acceptable multiple pronunciations (5 compared with 12 in the full version).

In short, IRT analysis of responses to the 50-item NART has produced a rational, defensible short form of the scale comprising only half the number of words yet still capable of accurately measuring the spectrum of verbal ability. We would thus recommend that we give serious consideration to inclusion of the short form of NART in the wave 12 main survey.[5]

Table 17: Proposed 25-item NART

Item #	*Item # in full NART*	*Word*	*% correct*
1	1	Chord	95
2	4	Aisle	95
3	10	Debt	94
4	14	Naive	88
5	5	Bouquet	84
6	35	Placebo	84
7	22	Subtle	82
8	25	Gouge	76
9	21	Hiatus	75
10	18	Heir	74
11	13	Equivocal	68
12	12	Rarefy	64
13	31	Facade	63
14	32	Zealot	55
15	26	Superfluous	52
16	30	Cellist	46
17	29	Quadruped	42
18	43	Leviathan	42
19	36	Abstemious	39
20	44	Beatify	28
21	46	Sidereal	28
22	41	Gauche	27
23	37	Detente	20
24	48	Syncope	12
25	47	Demesne	4

[5] A separate paper documenting in detail the development of this 25-item version of the NART is currently in preparation

Task Administration

General observation – Interviewers reported that the two tasks involving numbers were not so well received. By comparison, the other three tasks were seen as something novel, fresh and interesting, and in the case of Category Fluency and Symbol Digits, were even seen as "fun" by many.

Interviewer observations – Roy Morgan Research (RMR) recommended we collect additional data on whether task performance was adversely affected by disturbances (such as the presence of others, telephones, etc). We have, therefore, inserted the following additional interviewer observation for each task into the script for the main survey

INTERVIEWER RECORD: WAS PERFORMANCE ON THIS TASK ADVERSELY AFFECTED BY ANY SIGNIFICANT DISTRACTION OR DISTURBANCE [Yes / No]

Category Fluency

(i) Interviewers felt that it should be made clear to respondents that sub-categories of animals are permissible. We, however, prefer to follow the accepted preamble used for this test and so did not not endorse this suggestion.
(ii) RMR suggested that interviewers need more instruction about what is a valid response for the food category. This will need to be addressed in the manual (assuming this task is retained).
(iii) The CAPI script needs to be amended so that the category used in the task (animals or food) is identified at time of task administration (rather than later in the survey instrument).
(iv) The clock function in the CAPI console used to time this test may need to be fine-tuned to give interviewers a moment to get ready.
(v) More time during interviewer training may need to be devoted to how to record responses, including different types of short hand systems that might be used.
(vi) Pre-training sound files need to be amended to be more realistic (and less frightening).
(vii) The format of the booklet used to record responses may need to be reviewed with a view to making it more compact.

Backwards Digit Span

(i) RMR recommends that the preamble be amended to include the instruction: "Note that I cannot repeat the numbers after I have said them once." We endorsed this recommendation.
(ii) Some concerns were expressed about the variability in speed wuith which interviewers were n reading numbers by interviewers. RMR have suggested exploring using sound files delivered by CAPI for the main survey. We believe this is too risky without any pre-testing in field situations.

Number Series

(i) The preamble was regarded by interviewers as excessively long.

Symbol Digits Modalities

(i) Some interviewers reported difficulty in getting the respondent to stop at the 90 seconds mark. As a result, it is recommended that in the interviewer manual we add an instruction requesting interviewers to note the point at which the respondent got to at the 90 second mark, and only record the score based on entries up to that point.

(ii) There seems to have been some inconsistency in the extent to which interviewers checked the 10 practice examples. The need to do this will need to be reinforced in training.

(iii) On occasion, respondents looked at previous pages in the booklet used to record answers. More generally, we are concerned that the use of a booklet to record answers, rather than single loose pages (as was originally intended) could assist respondents (e.g., because previous answers may be visible through the back of the previous answer page or via indentations on the current answer page caused by the previous respondent). This suggests the need for this booklet to be redesigned. RMR is investigating the possibility of a page separator. Other options include thicker (card weight) paper or forms on detachable (tear-away) sheets.

NART

(i) The main issue reported by interviewers was that some respondents went very fast and interviewers were often reluctant to slow the respondent down, especially if the interview had already been going a long time. RMR have suggested revising the preamble to emphasise to respondents that they must wait until interviewer indicates it is okay before proceeding on to the next word, but we felt the existing preamble was more than adequate in making this clear.

(ii) Interviewers suggested that as part of the training materials, additional sound files with different speakers and accents be provided. We endorsed this proposal.

(iii) The script for the NART included an attempt at providing a phonetic guide to pronunciations usable by lay people. Interviewers seemed to regard this as a very useful aid, but nevertheless there are concerns that the way the written forms were interpreted could have confused some interviewers with respect to some words (most notably, "superfluous"). This guide will thus need to be revisited.

Summary

Category Fluency – This was the task that was best received by participants, but did not have the properties to suggest it is a good test of crystallized intelligence (as hoped). This was reflected in the pattern of scores across age cohorts. Scores may also be correlated with interviewer speed in recording answers. There are also difficulties arising from administering two forms of this test (animals and foods).

Backwards Digit – Seen as quite demanding by respondents, but was also quite short.

Number Series – Was the task that was least well received by participants, the distribution of scores was furthest from normal, and performance appears to vary with mode.

NART – The only measure of crystallized intelligence. Its main weaknesses are: (i) it is measuring something quite different among non-English speakers; (ii) it is the task that is most demanding on interviewers, which may result in poor inter-interviewer reliability; and (iii) it cannot be easily administered over the telephone. It is also quite long, but a much shorter version with reasonable properties could be administered in place of the full NART.

Symbol Digits Modalities – Reasonably well received by participants and least sensitive to both interviewer effects and learning behaviour by other household members yet to be interviewed. The instructions, however, are long and complicated. It also cannot be easily administered over the telephone.

Recommendation

Following the conclusion of the wave 12 DR we reported to the Department of Families, Housing, Community Services and Indigenous Affairs (FaHCSIA), and recommended the inclusion of the following three measures in the interview instruments to be used in the wave 12 main survey:

(i) Backwards Digit Span;
(ii) Symbol Digits Modalities; and
(iii) the short-form (25 item) version of NART

Telephone respondents will only be administered Backwards Digit Span.

This recommendation was subsequently endorsed by FaHCSIA.

References

Anger, S. & Heineck, G. (2008). *Do Smart Parents Raise Smart Children? The Intergenerational Transmission of Cognitive Abilities* (LASER Discussion Paper No. 23). Labor and Socio-Economic Research Centre, University of Erlangen-Nuremberg.

Australian Bureau of Statistics [ABS]. (2008). *2007 National Survey of Mental Health and Wellbeing: Summary of Result*s (ABS cat. No. 4326.0). ABS, Canberra.

Beardsall, L. & Brayne, C. (1990). Estimation of verbal intelligence in an elderly community: A prediction analysis using a shortened NART. *British Journal of Clinical Psychology* 29, 83-90.

Gray, M., D'Ardenne, J., Balarajan, M. & Uhrig, S.C.N. (2011). *Cognitive Testing of Wave 3 Understanding Society Questions* (Understanding Society Working Paper Series no. 2011-03). University of Essex: Colchester. [Available from: http://research.understandingsociety.org.uk/publications]

Jorm, A.F., Anstey, K.J., Christensen, H. & Rodgers, B. (2004). Gender differences in cognitive abilities: The mediating role of health state and health habits. *Intelligence* 32, 7-23.

Kiely, K.M., Luscz, M.A., Piguet, O., Christensen, H., Bennett, H. & Anstey, K.J. (2011). Functional equivalence of the National Adult Reading Test (NART) and Schonell reading tests and NART norms on the Dynamic Analyses to Optimise Ageing (DYNOPTA) project. *Journal of Clinical and Experimental Neuropsychology* 33(4), 410-421.

Lang, F.R., 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 Jahrbuch* 127, 183-192.

Lachman, M.E. & Tun, P.A. (2008). Cognitive testing in large-scale surveys: assessment by telephone. In S. Hofer and D. Alwin (eds), *Handbook on Cognitive Aging: Interdisciplinary Perspectives* (pp. 506-523). Sage: Thousand Oaks (CA).

Nelson, H.E. (1982), *National Adult Reading Test (NART) Test Manual*. NFER-Nelson: Windsor (UK).

Ofstedal, M.B., Fisher, G.G. and Herzog, A.R. (2005). *Documentation of Cognitive Functioning Measures in the Health and Retirement Study* (HRS Documentation Report DR-006). University of Michigan: Ann Arbor. [Available from: http://hrsonline.isr.umich.edu/index.php?p=userg]

Salthouse, T.A., & Prill, K.A. (1987). Inferences about age impairments in inferential reasoning. *Psychology and Aging* 2, 43-51.

Smith, A. (2007), *Symbol Digits Modalities Test: Manual* (10th printing). Western Psychological Press: Los Angeles.

Strauss, E., Sherman, E.M.S. & Spreen, O. (2006). *A Compendium of Neurospsychological Tests: Administration, Norms, and Commentary* (3rd ed.). Oxford University Press: Oxford.

Tombaugh, T.N., Kozak, J. & Rees, L. (1999). Normative data stratified by age and education for two measures of verbal fluency: FAS and animal naming. *Archives of Neuropsychology* 14, 167-177.

Watson, N. & Wooden, M. (2010), The HILDA Survey: Progress and future developments. *The Australian Economic Review* 43, 326-336.

Appendix A: The Measurement of Cognitive Skills in the HILDA Survey – A Proposal

by Bryan Rodgers (ANU) and Tim Windsor (Flinders University)

Background

Cognitive assessments are used for many purposes. Their practical utility includes selection, such as entry selection for educational courses or jobs, clinical identification of neuro-psychological deficits (e.g., after injury or from neurodegenerative disease), and a wide range of research purposes including the description of cognitive ageing over the lifespan, the prediction of individual differences in achievement and wellbeing, and the identification of environmental factors that enhance or diminish cognitive performance. Given this wide range of uses of cognitive tests it is unsurprising that many measures are available and that they differ hugely in their content, the psychological processes that underpin performance, modes of administration, and the expertise and resources required to derive measures from assessments of individuals.

It is not necessary to outline the above in any detail in considering what measures might be appropriate for inclusion in HILDA. However, there are some issues to document as the basis for decisions as to whether to include cognitive measures and, if so, which measure(s) to choose.

Purpose

The most likely use of cognitive measures in HILDA would be as additional indicators of an area of human capital that it not well covered in the existing data. Aside from educational history and qualifications, there is a shortage of information on the intellectual capacity of participants. Additional measures in this domain could be of value in predicting outcomes such as income, personal wellbeing, and demographic outcomes and transitions. A secondary use could be the assessment of change in cognitive performance over time if measures were repeated in a future wave. While this would not provide the depth of information normally collected in dedicated studies of cognitive ageing, it could provide some useful results given the breadth of risk factors included in the HILDA data set.

Content

The nature of the HILDA data collection imposes constraints on the content of test materials. Many population-based studies of cognitive abilities utilise computerised testing that can administer the types of tests that have traditionally been used in laboratory settings. However, pencil and paper versions of certain tests are also used with large-scale community samples and some studies have even utilised tests administered by telephone. A very important aspect of content for HILDA is that the study members warm to the nature of the assessment and do not find it threatening.

Underlying psychological processes

Historically, there have been different ways of conceptualising the "structure" of human intellectual abilities (typically informed by methods of factor analysis). For HILDA, we are not concerned with very specific capacities that map onto particular neural structures and networks as these are more relevant to clinical studies. The most common way to subdivide intellectual functions is the distinction between fluid and crystallised intelligence, attributed to Cattell or Cattell-Horn http://en.wikipedia.org/wiki/Fluid_and_crystallized_intelligence.

Previously, distinctions had been drawn between non-verbal and verbal intelligence, and also between abilities that were considered innate and those more affected by education. There are

threads that run through these conceptual frameworks and the measures used to assess particular abilities. Studies have also assessed psychomotor skills using a range of timed tasks such as reaction time (RT) measures. More complex RT tasks tap into higher-order abilities but the simpler tasks have relatively low correlations with other forms of cognitive testing. Given, HILDA's constraints for fieldwork, we can essentially ignore RT tasks.

Expertise required for administration

It is not feasible and it would be unnecessary to use measures in HILDA that required specific expertise in test administration. Tests are available that can be administered by survey interviewers with little additional training and such tests have been used previously in a range of settings and a variety of populations.

Shortlisting possible tests

Because of the wide-range of tests available, it would take a long time to document the relative merits of every one. Instead, we considered the specific requirements of HILDA and, initially independently and then following consultation and discussion, selected two tests representing each of the fluid and crystallised domains. These all had to have relatively short administration times, were easy to administer and to score, and were well used and respected by researchers in the fields of developmental psychology and cognitive ageing.

Measures of crystallised ability

Verbal Fluency is typically assessed by asking participants to name as many examples within a particular category (e.g. animals) in a specified time, usually 1 minute. Variations on the approach use a phonemic restraint (e.g. words beginning with a certain letter), rather than semantic, and it is possible to combine these (e.g. animals beginning with a certain letter). The task can be administered in several ways but the most likely for HILDA is for the interviewer to record verbal responses and score the total. (Some studies have also scored repetitions of responses, but this wouldn't be necessary for HILDA.)

Spot the Word is a task where real words and nonsense words are paired and the participant has to choose the real word. The words are presented in written form and the task gets progressively more difficult. The traditional form of the task uses 60 word pairs and so testing takes a relatively long time (Trish Jacomb, Manager of the PATH Through Life Project estimated about 4 minutes). Although this may rule out the task for HILDA, we are looking to see if shorter forms of the test have been used and whether reliability and validity is reasonable for a short form. This has proven difficult, probably because of the copyright restraints of test materials. Even tests that are free to use can place restrictions of not modifying the test materials.

Measures of fluid ability

Digit Span Backwards is used in traditional intelligence tests (including well used Wechsler scales). In the verbal forms of assessment, numbers (from 1 to 9) are read out at one-second intervals and the participant has to repeat the list in reverse order. Initially, the lists are short (usually starting with just two digits) and are lengthened progressively. The approach most likely for HILDA is to present two examples of each length (two digits, three digits etc) until the participant gets both examples wrong for a particular length. The score is the number of correct responses up to that point. The best estimate I have for this task is that it takes under 2 minutes (again thanks to Trish and PATH) but the length varies across participants.

The *Symbol Digits Modalities Test* requires participants to use a key which pairs up symbols with digits. The participant then works sequentially through a printed grid of symbols and writes the appropriate digit next to each symbol. The task is typically timed for 90 seconds

and the score is the number of correct digits written in that time. Again, this seems the type of task that might be shortened but we have not yet identified a published shorter version. Note that the administration time is increased above 90 seconds because the task must be explained to the participants. The test sheets can be scored later.

Recommendations

- We recommend piloting as many as possible of the four shortlisted measures.
- Ideally, one fluid measure and one crystallised measure would be used in Wave 12 but time constraints may preclude this possibility.
- We should explore further the possibilities of using short forms of some of the tests, particularly the Symbol Digit Modalities Test and Spot the Word

Table A1: Key Characteristics of the Four Selected Measures

Test	*F v C*	*Medium*	*Time*	*Age decline**	*Admin.*
Digit span backwards	F	Verbal	Up to 2 mins	Yes	Some scoring needed
Symbol digits modalities	F	Written	90 secs + explanation	Yes	Scored later
Verbal fluency	C	Verbal	1 min + explanation	Less so	Responses recorded
Spot the Word	C	Written	4 minutes	Less so	Scored later

* Crystallised tests can show improvements with age through to old age whereas fluid test can show declines at fairly young adult ages. It should be kept in mind that age differences in cross-sectional samples reflect cohort differences as well as longitudinal age changes. Any test measures would need to be used in conjunction with chronological age in HILDA.

Appendix B: Questionnaire Script, Wave 12 DR – Cognitive Ability Tasks

INTRODUCTION TO BE READ OUT:

A special feature of the of the interview this year is the next section, which comprises some exercises that involve you remembering and making judgments about words and numbers.

Your participation is completely voluntary.

As with all information you provide in this survey, the answers you give me will be confidential and used for statistical analysis only. I won't, therefore, be able to give you any specific feedback about your answers.

It will take about 12 minutes of your time.

O1 INTERVIEWER TO RECORD:

Continue with tasks 1
Discontinue tasks ..... 2 ➔T1

O2 *Have the record sheet ready to record the respondent's answers for this task. Have stop watch ready to count 60 seconds.*

I am going to give you a category and I want you to name things that belong in that category. Let's practice with the category "fruit". You could say peach or pear. Can you think of any other fruits?

(Wait for 2 correct items.)

In a moment I will give you another category. When I say "begin", I want you to name all the things from this new category you can think of, as fast as you can. You will have one minute to do this. I will let you know when your time is up.

The new category is animals. Do you have any questions? Ready?

If person stops before 1 minute is up, say: "There's still more time. Can you think of any more?"

If person asks whether birds, fish, insects, reptiles etc. are acceptable, say yes. If a participant says a category such as "bird", then names a specific type of bird (e.g., "sparrow"), then credit is given for each response.)

Names of mythical animals such as dragons and unicorns, and extinct animals (e.g., dinosaurs) are acceptable, but not given names of animals (e.g. Fido the dog).

INTERVIEWER TO RECORD: IS IT OKAY TO START THE TASK?

Yes, start task ..... 1 ➔O3
No, cannot understand instructions ..... 2 ➔O5
No, refused ..... 3 ➔O5

O3 INTERVIEWER SAY: **Begin.**

INTERVIEWER: AFTER THE 60 SECONDS IS UP, TALLY THE RESPONSES AND RECORD:

Total words	
Repeated words	
Intrusions	

O4a INTERVIEWER RECORD: WAS THE TASK COMPLETE IN FULL?

Yes ..... 1 ➔O4c
No ..... 2 ➔O4b

O4b INTERVIEWER RECORD: WHY DID THE TASK HAVE TO BE CUT SHORT?

MULTI RESP

Excessive distraction 1
Physical disability made completion impossible 2
Inability to understand the instructions 3
Extreme anxiety or discomfort 4
Refused to continue / doesn't want to do test 5
English language problems 6
Other (please specify) 7

O4c INTERVIEWER RECORD: WAS ANYONE ELSE PRESENT DURING THIS TASK? IF SO, WHO?

MULTI RESP

Yes, another sample member ..... 1 ➔O4d
Yes, child / children under 15 ..... 2 ➔O4d
Yes, non-household member ..... 3 ➔O4d
No, no one ..... 4 ➔O5

O4d INTERVIEWER RECORD: DID THIS PERSON HELP OR ASSIST THE RESPONDENT IN COMPLETING THE TASK?

Yes ..... 1 ➔O5
No ..... 2 ➔O5

O5 **I am now going to read out some lists of numbers, and I want you to repeat the numbers back to me in the reverse order from which I said them. So if I said "3, 8", you would say "8, 3". Do you understand?**

If respondent does not understand, repeat the instructions.

The sets will get larger as we go. And it may help if you close your eyes to help you concentrate.

INTERVIEWER TO RECORD: IS IT OKAY TO START THE TASK?

Yes, start task ..... 1 ➔O6
No, cannot understand instructions ..... 2 ➔O8
No, refused ..... 3 ➔O8

When the respondent gets one trial correct at a "level" move on to the first trial at the next level. If the first trial is incorrect, administer the second trial. If both responses at the same level are incorrect, the test is discontinued.

Read in monotone, 1 second per number. Drop your voice on the last digit to indicate it is time to respond.

If participant immediately self-corrects, do not count as an error.

If the participant asks for repetition, say: "I'm sorry, I can't repeat items."

O6 READ OUT: **OK, I'll start now.**

	READ OUT	*CORRECT ANSWER*	*CORRECT*	*INCORRECT*
1a	**2, 4**	(4, 2)	1	2
1b	**5, 7**	(7, 5)	1	2
2a	**6, 2, 9**	(9, 2, 6)	1	2
2b	**4, 1, 5**	(5, 1, 4)	1	2
3a	**3, 2, 7, 9**	(9, 7, 2, 3)	1	2
3b	**4, 9, 6, 8**	(8, 6, 9, 4)	1	2
4a	**1, 5, 2, 8, 6**	(6, 8, 2, 5, 1)	1	2
4b	**6, 1, 8, 4, 3**	(3, 4, 8, 1, 6)	1	2
5a	**5, 3, 9, 4, 1, 8**	(8, 1, 4, 9, 3, 5)	1	2
5b	**7, 2, 4, 8, 5, 6**	(6, 5, 8, 4, 2, 7)	1	2
6a	**8, 1, 2, 9, 3, 6, 5**	(5, 6, 3, 9, 2, 1, 8)	1	2
6b	**4, 7, 3, 9, 1, 2, 8**	(8, 2, 1, 9, 3, 7, 4)	1	2
7a	**9, 4, 3, 7, 6, 2, 5, 8**	(8, 5, 2, 6, 7, 3, 4, 9)	1	2
7b	**7, 2, 8, 1, 9, 6, 5, 3**	(3, 5, 6, 9, 1, 8, 2, 7)	1	2

Once both trials at the same level are incorrect, say:
Ok, that's all of those we need to do.

O7a INTERVIEWER RECORD: WAS THE TASK COMPLETE IN FULL?

Yes ..... 1 ➔O7c
No ..... 2 ➔O7b

O7b INTERVIEWER RECORD: WHY DID THE TASK HAVE TO BE CUT SHORT?

MULTI RESP

Excessive distraction 1
Physical disability made completion impossible 2
Inability to understand the instructions 3
Extreme anxiety or discomfort 4
Refused to continue / doesn't want to do test 5
English language problems 6
Other (please specify) 7

O7c INTERVIEWER RECORD: WAS ANYONE ELSE PRESENT DURING THIS TASK? IF SO, WHO?

MULTI RESP

Yes, another sample member ..... 1 ➔O7d

Yes, child / children under 15 ..... 2 ➔O7d

Yes, non-household member ..... 3 ➔O7d

No, no one ..... 4 ➔O8

O7d INTERVIEWER RECORD: DID THIS PERSON HELP OR ASSIST THE RESPONDENT IN COMPLETING THE TASK?

Yes ..... 1 ➔O8

No ..... 2 ➔O8

O8 INTERVIEWER READ OUT: **In this next task I will read you a series of numbers that may get larger or smaller in value. At the end you will try to figure out what the next number would be. So if the numbers were 2, 4, 6, 8, 10, the next number would be 12, as each number has increased by two.**

After I say each number I will pause for as long as you need, and then you should say "okay" when you are ready for me to go on. So if I said 2, you would say "okay" or nod when you are ready for me to go on to the next number. Then I say 4, and you say "okay" or nod, and so on.

Let's try one for practice.

35 … *(okay / nod)* **… 30 …** *(okay / nod)* **… 25 …** *(okay / nod)* **…20** *(okay / nod)* **… 15** *(okay / nod)* **AND the next number would be …?**

The answer should be 10 as each number has decreased by 5.

There will be different patterns, and some of these will be harder than others, so just do the best you can. If you are not sure of the answer, it is okay to guess. Do you have any questions?

If the respondent either asks to use a pen / pencil or picks up a pen, say "Please do not use a paper and pencil (pen) for any of these questions."

Interviewer, pause after each of the first 4 items for the "okay" response.

If participant immediately self-corrects and gets the right answer (e.g. "47 … no, 48"), record as giving correct answer.

If the participant asks for repetition, say: "I'm sorry, I can't repeat items."
INTERVIEWER TO RECORD: IS IT OKAY TO START THE TASK?

No, cannot understand instructions ..... 1 ➔ O9

No, cannot understand instructions ..... 2➔O11

No, refused ..... 3➔O11

O9 READ OUT: **OK, I'll start now.**

	READ OUT	*CORRECT ANSWER*	*CORRECT*	*INCORRECT*
N6a	**18 … 20 … 24 … 30 … 38 … AND the next number is?** **Okay. Are you ready for another? The next set is:**	(48)	1	2
N6b	**81 … 78 … 75 … 72 … 69 … AND the next number is?** **Okay. Are you ready for another? The next set is:**	(66)	1	2
N6c	**7 … 12 … 16 … 19 … 21 … AND the next number is?** **Okay. Are you ready for another? The next set is:**	(22)	1	2
N6d	**28 … 25 … 21 … 16 … 10 … AND the next number is?** **Okay; ready for the final one?**	(3)	1	2
N6e	**20 … 37 …18 … 38 … 16 … AND the next number is?**	(39)	1	2

O10a INTERVIEWER RECORD: WAS THE TASK COMPLETE IN FULL?

Yes ..... 1 ➔O10c
No ..... 2 ➔O10b

O10b INTERVIEWER RECORD: WHY DID THE TASK HAVE TO BE CUT SHORT?

MULTI RESP

Excessive distraction 1
Physical disability made completion impossible 2
Inability to understand the instructions 3
Extreme anxiety or discomfort 4
Refused to continue / doesn't want to do test 5
English language problems 6
Other (please specify) 7

O10c INTERVIEWER RECORD: WAS ANYONE ELSE PRESENT DURING THIS TASK? IF SO, WHO?

MULTI RESP

Yes, another sample member ..... 1 ➔O10d
Yes, child / children under 15 ..... 2 ➔O10d
Yes, non-household member ..... 3 ➔O10d
No, no one ..... 4 ➔O11

O10d INTERVIEWER RECORD: DID THIS PERSON HELP OR ASSIST THE RESPONDENT IN COMPLETING THE TASK?

Yes ..... 1 ➔O11
No ..... 2 ➔O11

O11 INTERVIEWER, INDICATE SHOWCARD A

INTERVIEWER READ OUT: **I want you to read slowly down this list of words. Start here (***indicate first word)* **and read the word out loud. After each word please wait until I say "next" before reading out the next word. I must warn you that there are many words that you probably won't recognize; in fact <u>most</u> people don't know them, so just have a guess at these.**

INTERVIEWER TO RECORD: IS IT OKAY TO START THE TASK?

Yes, start task..... 1➔O12
No, cannot understand instructions..... 2➔O14
No, refused..... 3➔O14

O12 Ok, go ahead with the first word.

		Acceptable pronunciations	*CORRECT*	*INCORRECT*
1	**CHORD**	kord	1	2
2	**ACHE**	ayk	1	2
3	**DEPOT**	deppo, dee-poe	1	2
4	**AISLE**	ile	1	2
5	**BOUQUET**	boo-kay, boe-kay	1	2
6	**PSALM**	sarm, solm	1	2
7	**CAPON**	kay-pn, kay-pon	1	2
8	**DENY**	di-ny	1	2
9	**NAUSEA**	norz-ee-a, norse-ee-a, nor-ja	1	2
10	**DEBT**	det	1	2
11	**COURTEOUS**	kurt-ee-us	1	2
12	**RAREFY**	rare-i-fy	1	2
13	**EQUIVOCAL**	e-kwiv-e-kl, i-kwiv-e-kl, ee-kwiv-e-kl	1	2
14	**NAÏVE**	ny-eev	1	2
15	**CATACOMB**	kat-a-koam, kat-a-koom	1	2
16	**GAOLED**	Jayld	1	2
17	**THYME**	time	1	2
18	**HEIR**	air	1	2
19	**RADIX**	ray-diks	1	2
20	**ASSIGNATE**	ass-ig-nayt	1	2
21	**HIATUS**	hy-ay-tiss	1	2
22	**SUBTLE**	sut-l	1	2
23	**PROCREATE**	proe-kree-ayt	1	2
24	**GIST**	jist	1	2
25	**GOUGE**	gowj {"ow" as in "c<u>ow</u>"}	1	2

INTERVIEWER SAY: **OK, now go onto SHOWCARD B and continue.**

		Acceptable pronunciations	*CORRECT*	*INCORRECT*
26	**SUPERFLUOUS**	soo-pur-floo-ess,sa-pur-floo-ess	1	2
27	**SIMILE**	sim-i-lee	1	2
28	**BANAL**	ba-narl, ba-nal	1	2
29	**QUADRUPED**	kwod-roo-ped	1	2
30	**CELLIST**	chel-ist	1	2
31	**FACADE**	fa-sard, fass-ard	1	2
32	**ZEALOT**	zel-it	1	2
33	**DRACHM**	dram	1	2
34	**AEON**	ee-on	1	2
35	**PLACEBO**	ple-see-boe	1	2
36	**ABSTEMIOUS**	ab-stee-mee-us	1	2
37	**DETENTE**	day-tont	1	2
38	**IDYLL**	idol idd-il	1	2
39	**PUERPERAL**	pew-ur-per-el	1	2
40	**AVER**	a-vur	1	2
41	**GAUCHE**	goe-sh	1	2
42	**TOPIARY**	toe-pee-e-ree, toe-pee-err-ee	1	2
43	**LEVIATHAN**	le-vy-e-then	1	2
44	**BEATIFY**	bee-at-i-fy	1	2
45	**PRELATE**	prel-it	1	2
46	**SIDEREAL**	sy-dear-ee-el	1	2
47	**DEMENSE**	di-mayn, di-meen	1	2
48	**SYNCOPE**	sink-e-pee	1	2
49	**LABILE**	lay-bile	1	2
50	**CAMPANILE**	kam-pe-nee-lee, kam-pe-nee-lay, kam-pe-neel	1	2

O13a INTERVIEWER RECORD: WAS THE TASK COMPLETE IN FULL?

Yes ..... 1 ➔O13c
No ..... 2 ➔O13b

O13b INTERVIEWER RECORD: WHY DID THE TASK HAVE TO BE CUT SHORT?

MULTI RESP

Excessive distraction 1
Physical disability made completion impossible 2
Inability to understand the instructions 3
Extreme anxiety or discomfort 4
Refused to continue / doesn't want to do test 5
English language problems 6
Other (please specify) 7

O13c INTERVIEWER RECORD: WAS ANYONE ELSE PRESENT DURING THIS TASK? IF SO, WHO?

MULTI RESP

Yes, another sample member ..... 1 ➔O13d
Yes, child / children under 15 ..... 2 ➔O13d
Yes, non-household member ..... 3 ➔O13d
No, no one ..... 4 ➔O14

O13d INTERVIEWER RECORD: DID THIS PERSON HELP OR ASSIST THE RESPONDENT IN COMPLETING THE TASK?

Yes ..... 1 ➔O14
No ..... 2 ➔O14

O14 *Have card to hand to respondent and stop-watch ready. Stop the exercise after 90 seconds.*

This next exercise involves matching numbers to symbols.

(Hand self complete card to participant)

Please look at the key at the top of the page. The symbol in the top row matches the number in the box below it.

Now look at the next line of boxes (*interviewer point to the line of boxes***). Notice that the boxes below the symbols are empty. Your task is to fill each empty box with the number that matches the symbol using the key at the top of the page. Is that clear?**

(If respondent requires further instructions say): Please look again at the key on top of the page, each of these symbols in the top row has a matching number. Your task is to fill in the blank boxes underneath each symbol (point) using the key at the top of the page to match the number? Is this clear?

Let's have a go at the first symbol. Looking at the key, you will see that number 1 goes in the first box, so write the number 1 in the first box. Now what number should you put in the second box? (*Number 5***) That's right. So write the number 5 in the second box. What number goes in the third box? (***Number 2***) Two, right.**

For practice, fill in the remaining boxes and stop at the double line.

Interviewer check practice boxes. Any errors made in these practice responses should be immediately pointed out. If needed, you will need to explain the task again.

Now when I say "Go!" write the numbers just like you have been doing until I say "Stop!", starting from here (*interviewer point to the first box after the double line***). When you come to the end of the first line, go quickly to the next line without stopping. If you make a mistake, just write the correct answer over your mistake. Don't skip any boxes and work as quickly as you can.**

Any questions?

INTERVIEWER TO RECORD: IS IT OKAY TO START THE TASK?

Yes, start task ..... 1 ➔O15
No, cannot understand instructions ..... 2 ➔T1
No, refused ..... 3 ➔T1

INTERVIEWER SAY:
OK, begin

O15 INTERVIEWER: AFTER THE 90 SECONDS IS UP, TALLY RESPONSES AND RECORD:

Correct responses ☐

O15a INTERVIEWER RECORD: WAS THE TASK COMPLETE IN FULL?

Yes ..... 1 ➔ O15c
No ..... 2 ➔ O15b

O15b INTERVIEWER RECORD: WHY DID THE TASK HAVE TO BE CUT SHORT?

MULTI RESP

Excessive distraction 1
Physical disability made completion impossible 2
Inability to understand the instructions 3
Extreme anxiety or discomfort 4
Refused to continue / doesn't want to do test 5
English language problems 6
Other (please specify) 7

O15c INTERVIEWER RECORD: WAS ANYONE ELSE PRESENT DURING THIS TASK? IF SO, WHO?

MULTI RESP

Yes, another sample member ..... 1 ➔O15d
Yes, child / children under 15 ..... 2 ➔O15d
Yes, non-household member ..... 3 ➔O15d
No, no one ..... 4 ➔T1

O15d INTERVIEWER RECORD: DID THIS PERSON HELP OR ASSIST THE RESPONDENT IN COMPLETING THE TASK?

Yes ..... 1 ➔T1
No ..... 2 ➔T1