Version - July 2009

SCOTTISH SURVEY OF ADULT LITERACIES 2009: TECHNICAL REPORT

SCOTTISH SURVEY OF ADULT LITERACIES 2009: TECHNICAL REPORT

Edited by:

Ralf St.Clair
University of Glasgow

2010

Contents

1. INTRODUCTION

This technical report is the companion to the SSAL2009 Report of Findings and SSAL2009 Research Findings documents available from the Scottish Government. These two documents discuss the findings of the survey and their meaning, therefore these topics will not be addressed in the current document. The purpose of this technical report is to present the details of the conduct of the research from fieldwork through to the production of the findings.

As the first national survey of literacy and numeracy abilities to be conducted in Scotland in 13 years, the outcomes of SSAL2009 have profound implications for policy and practice. As a small nation with a strong commitment to social equity and inclusion, Scottish Government decision-making requires reliable and sophisticated information. This report summarises key information that will enable decision makers to familiarise themselves with the study's key characteristics and the methodology applied.

This report is presented in four main chapters. Chapter 2 provides an overview of the multipartner project and the overall management of the project. The following three chapters reflect the three phases of the empirical research.

Chapter 3 discusses the fieldwork phase, including the extensive quality control procedures at this stage of the work. The work was undertaken, and the chapter written, by Zsofia Bohar of Gallup Europe.

Chapter 4 discusses how the data from the fieldwork was scored and scaled—in other words how raw data was analysed to produce information on the literacy scale placement of the Scottish population. This work was conducted by the Educational Testing Service (ETS), based in Princeton, New Jersey, USA, and the chapter was written by Kentaro Yamamoto, the team leader on this project.

Chapter 5, written by the National Foundation for Educational Research's (NFER) lead statistician Dougal Hutchinson, discusses some of the key issues around validity that are faced when linking background data to literacy scores produced from an Item Response Theory model.

The appendices include further detail on the fieldwork, the analysis and the instruments used. Appendix A shows the progress of interviewing over the fieldwork period. Appendix B shows the percentage of correct responses for each item in the cognitive task booklets as a unweighted and weighted figure. Appendix C has the same information but grouped according to the test block of the items.

Appendix D shows the social variables used in the survey, such as the educational career of individuals, their social class, deprivation of the areas surveyed and whether they could be considered as urban or rural.

Appendix E contains the entire text of the background questionnaire (with most interviewer instructions removed) and Appendix F contains four sample stimuli and associated questions.

2. THE CONDUCT OF SSAL2009

Ralf St.Clair
University of Glasgow

This chapter provides an overview of the way the project was organised, focusing mainly upon the management of the consortium and quality assurance. More details of the specific techniques can be found in the following chapters. It was a complex project spanning—if the period from tender to final report is included—two years of work for most of the partners. The consortium was led by Ralf St.Clair, Kathy Maclachlan and Lyn Tett, established literacy researchers from the Universities of Glasgow and Edinburgh.

There were three partner organisations:

- Gallup Europe were responsible for the actual data collection interviews through their UK based partners Blauw. They also collated the background questionnaire data.

- Education Testing Services took the literacy test data booklets, scored them, and put the scores into levels using the techniques developed through years of work on international literacy surveys including IALS in the 1990s.

- National Foundation for Educational Research linked the literacy and social background data together so that it was possible to understand how social factors affect literacy levels.

Each of the partners had specific responsibilities in the project. The consortium was committed to delivering robust information to the Scottish Government, and worked closely with individuals from Lifelong Learning Analytical Services Unit in the Education Analytical Services Division throughout the survey development, delivery and analysis process.

There were five main stages to the project (Table 1). The first was to develop instruments. As the last time this survey had been used was 1996, the computer files available to the research team were not usable. Since the background questionnaire, an extremely detailed set of questions on work, education and life, was delivered through computer last time this was a significant problem. Instruments had to be recreated from the answer categories available from ETS, a paper copy of the old questionnaire that did not include question wording, and an English Canadian version of the survey that did include actual wording.

It took a considerable amount of time to recreate the instruments. Due to this delay it was decided to implement a pen and paper background questionnaire instruments rather than spend more time developing computer based (CAPI – computer assisted personal interviewing) versions of the survey.

There was a similar issue with the literacy test booklets, where initially only paper copies were available, making it hard to reproduce them in sufficient quality. NFER

recreated some of the graphics and text, while in other cases original graphics were able to be found. It proved possible to create a set of booklets of usable quality by combining sources.

ETS approved the final version of the survey instruments and the management plan for moving data from pen and paper instruments to database. Given their experience in the field of literacy surveys their input was appreciated.

Table 1: Stages of SSAL2009

Central task	Lead partner	Quality assurance
Preparation of Instruments	Universities/All	The instruments were checked and approved by ETS and NFER before use.
Data collection/ fieldwork	Gallup and Blauw	There is a full report on the QA procedures used in chapter 3.
Scoring and scaling of literacy assessments	Education Testing Service	Description of the model used can be found in chapter 4. ETS have several decades of experience with this type of data.
Matching scores to background data	National Foundation for Educational Research	Description of some of the challenges faced during this process can be found in chapter 5.
Preparation of findings and reports	Universities	Reports were developed in response to Scottish Government requirements, and went through a number of iterations.

The collection of data was a door to door survey conducted in Spring and Summer of 2009. Gallup were the leaders on this part of the project, and they have provided a full description of their approach in chapter 3. Fieldwork staff also handled the data entry for the background questionnaire, which was entered immediately into an electronic database using a standard ETS data collection framework. While the administration of this size of sample is never straightforward, Gallup produced a very high quality dataset and were scrupulous regarding quality controls.

In July 2009 the data were transferred to NFER. They copied every page from the literacy tests for security and sent the originals to ETS in Princeton. This constituted around 50,000 pages overall. NFER also supplied the weighting variables for ETS to use in analysis of the responses.

ETS scored the tests as part of their contract with the Scottish Government. Further information on this process and the levelling procedures can be found in chapter 4. To ensure the quality of the scoring, 10% of the scores were checked with NFER, who were able to double check by comparison with the security copies of the completed instruments.

In November 2009 main responsibility for the data analysis shifted back to NFER. They were responsible for the descriptive statistics in terms of who responded to the survey and the analytical statistics investigating the links between the social characteristics captured in the background questionnaire and literacy scores.

NFER's work fell into two main components. One was development of crosstabs showing how different parts of the population would score based on the actual tested scores of the sample, including derivation of standard errors. There was also some testing for the significance of these relationships. The second component was a limited number of regression calculations to investigate which factors could predict literacy scores. For these calculations literacy scores were treated as continuous variables rather than divided up into levels. The challenges of matching literacy scores to social characteristics are described by NFER in chapter 5.

Throughout the data analysis processes figures and findings were extensively cross-checked to ensure that no computational errors had been made and that there was no inaccurate data processing. Where necessary specialists with knowledge of a particular area were asked to review and comment upon findings. For example, Rob van Krieken of the Scottish Qualifications Authority (SQA) was able to comment upon the use of the Scottish Credit and Qualifications Framework, as well as provide comments based on his experience with survey research.

The academic staff from the Universities of Glasgow and Edinburgh were responsible for producing the final report and the research findings document between March and July 2010. The reports were checked over both by the Scottish Government Educational Analytical Services Division and NFER, to ensure that no errors of interpretation went uncorrected.

The Research Advisory Group (RAG) played an important role throughout the project, providing insights from related policy areas including the teams that conduct the Annual Population Survey (APS), the Labour Force Survey (LFS) and so on. This helped a great deal with the development of measures and their interpretation.

In closing this chapter it is worth commenting upon the comparability of SSAL2009 with the IALS survey of 1996. In brief, comparisons should be avoided. Even though the same instrument was used, there were three very important differences. Firstly, the sample size was much larger in 2009 than in 1996 (1922 vs. 704). Secondly, geographical coverage was quite different. In 2009 we sampled across Scotland whereas in 1996 we did not sample North of the Caledonian Canal. Finally, during data analysis ETS suggested we modified the Item Response model for greater power. We did so, and while the new model is closer to that used in 1996 for international purposes, it is quite different from the 1996 UK model. Taken together, these changes mean that longitudinal comparisons are not viable.

One notable feature of the project was the amount of time and effort put into quality assurance procedures. With several partners involved in a complex and unusual project both collecting demographic information and asking people to undertake testing, there were many possibilities for difficulties. From the beginning of the project, the partners and the RAG paid conscious attention to the need to create clear and auditable pathways for the data. Data was shared between the partners, and procedures were made as transparent as possible.

Scottish Government personnel also called on the resources available within the government to enhance quality. A Senior Statistician was involved in quality control, and several researchers involved in other large-scale surveys were asked for guidance and to approve the quality control mechanisms. These individuals were satisfied with the procedures and the dataset of SSAL2009.

At the end of their work both Gallup Europe and ETS were asked to provide statements that the data was of appropriate quality, which they did. The researchers are highly confident that the survey findings are based upon the highest possible quality of data and analysis.

3. FIELDWORK

Zsofia Bohar
Gallup Europe

This chapter describes the Scottish Survey of Adult Literacies (SSAL2009) preparations and procedures in the field, including key quality indicators and assessment of the survey operations.

The SSAL2009 fieldwork was carried out via Blauw UK (as fieldwork partner), under the supervision of Gallup Europe. The SSAL2009 fieldwork and sampling was face-to-face. Preparatory activities for the study started in January 2009, and the fieldwork was launched on the 10th of March 2009 and closed at the end of June 2009. SSAL2009 collected 1,953 interviews and tests.

The sample of SSAL2009 covered the full geographical area of Scotland – including the Highlands and the Islands. SSAL2009 is representative of the Scottish resident population between 16 and 65 years-of-age living in private households. The sample is capable of producing robust results for certain subgroups of interest (e.g. gender, broad social class, occupation, educational background and level of urbanisation).

Overview of SSAL2009

The survey was conducted by personal interview in respondents' homes and consisted of two main elements, a background questionnaire and a literacy assessment. The background questionnaire collected information on the socio-demographic characteristics of the respondent such as age, sex, education, occupation and income as well as asking about literacy activities such as reading as part of their job or for pleasure, television viewing, and participation in training or adult education.

After taking part in the interview, respondents were asked to complete a short screening assessment which sought to identify those with very limited literacy skills. Respondents who correctly answered at least two of the six screening tasks were then asked to complete a larger assessment booklet which measured three literacy scales. These were:

Prose literacy The knowledge and skills required to understand and to use information from texts such as newspaper articles and passages of fiction.

Document literacy The knowledge and skills required to locate and to use information contained in various formats such as timetables, graphs, charts and forms.

Quantitative literacy The knowledge and skills required to apply arithmetic operations, either alone or sequentially, to numbers embedded in printed materials.

In order to ensure as broad a range of item content as possible the total number of tasks in the assessment was larger than any one individual could reasonably be asked to complete. Each respondent therefore was only asked to complete a subset of the total assessment. The assessment items were grouped into seven blocks and the blocks were arranged in different combinations into seven booklets. Each booklet contained three blocks of items and each block appeared at each possible location, the beginning, middle or end of a booklet in a spiral effect. Respondents were allowed to take as much time as they required completing the booklet.

Sampling

The issued sample for the survey consisted of approximately 7000 post code address file (PAF) addresses allocated proportionally in Scotland. The 7000 addresses were grouped in clusters of 28 each. Each of these belonged to a specified data zone[1], which is a statistical unit (composed by adjunct census output areas) comprising of about 500-1000 households.

The SSAL2009 sample did not use any explicit stratification. However, to ensure proper coverage an implicit stratification was implemented, by selection from a list of DataZones ordered by some key characteristics based on information available from the Scottish Neighbourhood Statistics data. Prior to systematic random selection of the PSUs (primary sampling units), all DataZones in Scotland were sorted by:

- Region
- Level of urbanisation (urban core / urban periphery / rural)
- Socio economic characteristics (Scottish Index of Multiple Deprivation, quintiles)

There were 250 such primary groups (clusters) selected as PSUs, using a fixed interval selection with a random starting point. This stage of sampling was carried out by Gallup. In each of the sampled PSUs 28 addresses were selected randomly (by a UK sample provider CACI), and issued for the study overall. Out of the 28 addresses the survey anticipated about 10 completes on average in each PSU. 50 replacement PSUs were also identified to substitute for primary PSUs that proved to be ineffective (e.g. very small DataZones that were depopulated since the enumeration, or where residents were totally non-cooperative).

In total, interviewers visited 5455 addresses and used 233 main and 3 replacement PSUs. Out of the 5455 addresses visited a total of 1953 interviews were achieved between March 10th and June 29th 2009.

The vast majority of the sample addresses were accurate with very few errors regarding multiple or single residence. An acceptable level of un-occupied/vacant properties was found and very few were found to be used as commercial sites (see tables 3 and 4 for details).

[1] http://www.sns.gov.uk/Guide/GeographyGuide.asp?GeographyType=ZN#Meta

Three PSUs were replaced as all original points were found to be unproductive. One area was extremely affluent and no respondents wished to take part in the study. The second and the third areas were generally found to have very few properties occupied (or answering the door).

Due to a late[2], and also unexpectedly slow start (see Appendix A. for information on the fieldwork progress), a decision had to be made regarding sample size. The choice was either to continue fieldwork longer to attempt a sample size of 2,500, or to keep to timetable and accept a smaller sample. It was decided through consultation with the Scottish Government that the latter strategy was necessary in order to ensure results by the end of the year, and the research team expressed the hope that Blauw (the agency responsible for data collection) could attain 2,000 responses. In reality 1,953 interviews were completed.

Interviewer selection and training

Interviewers with considerable face-to-face survey experience were recruited to conduct the SSAL2009 fieldwork. The interviewers for this study were selected on the basis of experience, with a minimum requirement of having been involved in at least 5 face–to-face fieldworks prior to SSAL2009. Overall, due to retention problems, over 90 interviewers worked on SSAL2009 during the whole fieldwork period.

Gallup developed a project-specific SSAL2009 Interviewing Guide that provided detailed documentation on address treatment, visiting rules, household selection at the address, respondent selection, interview and test administration. (In addition, a shorter brochure called SSAL2009 Golden Rules was issued to interviewers, drawing attention to key problem areas identified at the early checks of the survey implementation, see the section on quality control.)

All participating interviewers attended in-person training where their duties and the components of the SSAL2009 fieldwork administration were explained in detail. As part of the training, they had an opportunity to do an interview with a real-life respondent before starting the actual data collection. Training materials were provided to all interviewers to use as reference material. The main interviewer briefing took place on 4th March at Glasgow University from 9-4pm. An additional briefing took place in Aberdeen the next day on 5th March from 9-4pm.

Interviewers who could not make either briefing were trained in small groups or individually by the Supervisor or her deputy on a face to face basis.

[2] According to the original timing, 21 weeks were allocated to complete 2,500 interviews, but due to the delay with the finalisation of the fieldwork materials and also to the tight time allocation for the copying and shipping of the booklets a net of 16 weeks remained for fieldwork. The eventual sample size achieved was proportional to the decreased time window available for the assessment.

Table 2. Number of interviewers/supervisors per training session

Location of the Training	Number of participants	Training was held by:
Glasgow	48	Jenny Bradshaw (NFER) Henk Scholte (Blauw Director), Sheena Sidhu (Blauw Project Manager), Mark Luckraft (Field Director)
Aberdeen	12	Mark Luckraft and Jacqui Odowd (Supervisor) and Christine McNeally (Supervisor)
Other sessions (in small groups or individually)	38	Jacqui O'Dowd (Supervisor), Christine McNeally (Supervisor)

The local fieldwork supervision team conducted several reminder sessions, which were all done in-person by the two field supervisors responsible for the fieldwork execution in Scotland. Some of the interviewers considered the training session as far too long compared to what they were used to and a few of them chose to withdraw after they received training due to the complexity of the administration tasks.

Quality control

The survey had a quality control scheme matching to general ESOMAR guidelines and general practice.[3] As a minimum, a random 10% of all conducted interviews were to be verified. During the quality control process a multi-layered approach was applied by the local fieldwork company and Gallup Europe.

Gallup quality control activities

Checking of physical survey documents:

Gallup requested the first 2 completed interview materials from each interviewer working on the SSAL2009 to be scanned and sent to Gallup for checking purposes. As a result Gallup Europe checked a total of 82 interviews from 31 interviewers within the first 5 weeks of fieldwork.

Checking was based on the following aspects:

1. ID treatment (checking consistency of the PAF key, Gallup ID, Interviewer ID, Date, sticker on each survey document)
2. Usage of Kish-grid (correct respondent selection)

[3] http://www.esomar.org/

3. Booklet rotation (checking whether interviewer followed the rotation sheet when distributing the booklets)
4. Time recordings (checking the order of asking the survey materials by looking at start-end times of the questionnaires and the Main Task Booklets, checking too short/long timings)
5. Consistency of respondent's age, gender in the Background questionnaire vs. the Survey Administration Sheet
6. Checking if all seven survey documents are completed (including the Mock Newspaper)
7. Skipping (checking if skipping instructions were followed in the Background questionnaire)
8. Consistency check in the questionnaire
9. Probing (checking if interviewers did probe on the open ended questions e.g. D9a)
10. Fieldwork documentation (checking in both Booklets if interviewers understood and administered the attempts properly)
11. Scoring in the Core Task Booklet (checking if interviewers scored correctly and counted the right answers properly)
12. Accuracy of Main Task Booklet used (checking the type of Main Task Booklet vs. the rotation sheet and that the code circled in Q14 (Core Task Booklet) is identical with the booklet number and that it is copied into the Survey Administration Sheet correctly)
13. Outcome codes (checking if the final outcome code (P8) was assessed properly)
14. Handwriting in the Core Task Booklet vs. the Main Task Booklet

After checking the first 40 cases the following quality control issues were identified and addressed:

- Some of the interviewers were not precise when copying the IDs onto each survey document and they forgot to put the sticker on the Survey Administration Sheet.
- There were two instances when we found problems with respondent selection (ineligible person was included in the Kish-grid)
- 2 interviewers started the interview with the Booklets instead of the Background questionnaire.
- There were cases where the administration of the booklets was not fully correct; most of the problems concerned the assessment of the outcome codes (P8 and P9, describing results of the attempts and reasons for refusals) and the understanding of 'attempt'.
- Gallup has also found some minor problems in the background questionnaires (mainly wrong skipping).

In order to address these issues Gallup suggested issuing an information sheet ('SSAL2009 Golden Rules') to the interviewers reinforcing the most important elements. Blauw distributed this sheet to interviewers and the supervisors reported positive feedback on the sheet claiming that the interviewers understood the process and the need for accuracy. Gallup did another round of checking (further 42 cases) to make sure that the quality of interviewing had improved, which was indeed the case.

Interview verification by phone

Gallup back-checked 295 randomly selected interviews. 17% of the cases could not be recalled due to missing phone numbers (either because respondent refused to give a number or did not have a phone). About 4% refused to answer our back-checking questionnaire.

87% of the respondents confirmed that the interview took place with the right person (matching the Kish grid selection) and in the proper way. In a very few cases (1%) respondents claimed to have filled in the Booklets first and answered the background questionnaire afterwards.

Some of the respondents remembered a longer duration of the interview than the interviewer recorded on the survey documents. Sometimes the age of respondent did not match exactly with the one in the background questionnaire; however the difference was only 1 year.

Table 3. Gallup interview verification, results overview

Back-checking results	Calls	% of total called	% of total answered
A. Conducted properly	205	69	87
B. Order of asking the survey elements was wrong	6	2	3
B. Wrongly recalled the length of interview	15	5	6
C. Age did not match (mostly 1 year difference)	9	3	4
Total who answered the back-check interview (A+B+C)	**235**	**80**	**100**
D. Refused back-check interview	21	7	
E. Wrong number	39	13	
Total called (A+B+C+D+E+F)	**295**	**100 (rounded figure)**	

Although the back-checking results did not reveal any major problem, in week 17 Gallup visited the fieldwork team in Scotland to debrief the interviewers and get a direct feedback from them on the interviewing and all the difficulties they were facing. Supervisors were debriefed on the results of Gallup's back-checking and they again contacted the interviewers and reiterated to them the importance of asking the survey

elements in the right order. Gallup accompanied two interviewers to see how the actual interviewing was going and to understand the level of difficulty of the project by observing the field visits.

Data consistency checking

Besides providing metadata on the requested format the survey result were supposed to be recorded electronically (practically a blank SPSS datafile), Gallup created SPSS syntax files that aimed to uncover coding inconsistencies (e.g. logical verifications, and in some cases interval checks – e.g., for age, income, etc.) both for the survey datafile and the survey administration sheet data. Blauw was requested to correct/explain the discrepancies it discovered.

Blauw quality control activities

Assisted interviewing

Blauw completed approximately 30 supervisor/deputy accompanied visits. Any point where a questionnaire was found to have missing or questionable data was re-contacted by either telephone or in person by the supervisor to verify the information given. At any point where the questionnaires were not completed to a satisfactory standard by the interviewer, a supervisor accompanied them for a few interviews and the interviewer was re-briefed to ensure they understood all the technical requirements.

Respondent re-visits

Blauw's Quality Control Manager randomly selected 10% of the interviews and made three contact attempts for each individual respondent. The manager first checked that the respondent fit the age criteria and then asked some set questions about the actual interview, how long it took, whether the interviewer was present in the household throughout the whole survey. The back-checking calls also included a few questions regarding the interviewer, i.e. did they show their ID card, did they use show cards during the interview, etc.

Finally, three of the survey questions were repeated to ensure that the interview actually took place in its entirety (one question on the background questionnaire, one on the Core Task Booklet and one on the Main Task Booklet). Here the primary aim was to confirm the validity of the interview (that is, it indeed took place). If a fake interview was identified, it had to be replaced and all interviews belonging to the same interviewer had to be verified.

Obviously, proper briefing of the interviewers was essential in order to avoid future problems with correction of mistakes that would have required much more effort (see previous section).

Survey instruments

Overall, respondents found the survey interesting and were positive about helping the Scottish Government to measure literacy and numeracy skills. Interviewers did not report any major problems related to the background questionnaire or monitoring the attempts during the self-completion part of the interview.

Interviewers had difficulty only when assessing the outcome code for P8 (interview outcome, especially in terms of separating break-offs and refusals at the various stages of the interview) in the administration sheet of the Main Task Booklet, otherwise they managed to follow the clear guidelines given in the interviewer manual. The questionnaire was found to be straightforward, only a few interviewers had problems with skipping in section *D. Socio-demographics* especially when recording occupation for those who did not work in the past 12 months.

The duration of the interviews proved to be longer than anticipated, despite a fair amount of trimming of the background questionnaire during instrument preparation. The original IALS documents suggested a mean duration of 69 minutes, and interviewers had been telling respondents that they would last "about an hour." In reality interviewers were reporting that the assessment booklets *alone* took more than 50 minutes to complete in many instances resulting in the interview's total duration of 90-100 minutes. In the early stages of the fieldwork the length resulted in a higher than expected level of refusals and in some break-offs (especially by respondents declining to continue with or failing to complete the test booklet once the survey interview segment – Background Questionnaire – was over). Due to the low response rate, a cash incentive was necessary for completion of the task booklets.

Fieldwork outcomes

The overall response rate measures the proportion of persons interviewed out of all eligible persons sampled.

$$R = \frac{\text{Number of completed + partially completed interviews}}{\text{All addresses visited – wrong addresses (eligibility unknown) - other non-contact – ineligible households}}$$

R=0.59

The 59% actual response rate is in line with what we anticipated in the tender document (60%).

Technical data on the Survey Administration Sheets are detailed in the table below.

Table 4. Technical details of the Survey Administration Sheets

Number of target interviews:	2,000
Number of sampling points (PSUs) used: Main:	233
Number of sampling points (PSUs) used: Replacement:	3
Number of addresses visited in total:	**5,455**
Address not traceable	14
Vacant/uninhabited house/flat	38
Non-residence (e.g., business, institution)	6
Other non-contact	1,152
Number of households contacted in total	**4,245**
Number of households with no eligible respondent	910
Number of eligible households	**3,335**
Number of household gatekeeper (not the respondent) refuses cooperation (HARD refusal)	604
Number of households not available/not capable	54
Other (contact at household level)	14
Number of eligible respondent refuses cooperation (HARD refusal)	635
Number of eligible respondent not available (is in hospital, prison, travelling etc.)	49
Number of eligible respondent physically or mentally unable to participate	12
Other (contact at respondent level)	13
Number of terminated interviews	1
Number of fully completed interviews	**1,953**
Response rate	**59%**
Number of attempts in total (all 3 visits)	9,405
1. Number of completes by the 1st contact	1,052
2. Number of completes by the 2nd contact	538
3. Number of completes by the 3rd or more contact	363
A. Average lengths of the background questionnaire (min)	25
B. Maximum length of the background questionnaire (min)	95
C. Minimum length of the background questionnaire (min)	14
D. Total interviewing (background questionnaire only) time (hours)	814
E. Number of interviewers working on the project (in total)	91
F. Average number of contacts per household	1.7

Overall, refusal was mainly driven by general reluctance to be bothered while at home, scepticism towards surveys, confidentiality and the survey matter and not by respondents' capabilities for answering the questions. The most frequently (39%) recorded reason was that potential respondents did not want to be bothered by the interviewers. Almost a quarter of potential respondents refused to participate because they 'do not believe in surveys'. Reasons related to physical or

mental/cognitive capabilities concerned only 6% of the respondents in total (including those having language difficulties).

Table 5. Reasons for refusal/non-response

	N	%
1 Doesn't believe in surveys	312	23.9
2 Anti-government	22	1.7
3 Invasion of privacy	57	4.4
4 Concerns about confidentiality	42	3.2
5 Can't be bothered	514	39.4
6 Previous bad survey experience	35	2.7
7 Disliked survey matter	64	4.9
8 Death or unusual circumstance	8	0.6
9 Sickness or illness in household	59	4.5
10 Language difficulties	46	3.5
11 Reading and writing difficulties	11	0.8
12 Learning disability	5	0.4
13 Mental/emotional condition	1	0.1
15 Hearing impairment	3	0.2
16 Blindness/visual impairment	3	0.2
17 Speech impairment	1	0.1
18 Physical disability	3	0.2
20 Respondent is not capable (undefined)	8	0.6
21 Other	110	8.4
Total	1305	100

Across the project as a whole the vast majority (93%) of the interviews were complete, meaning that respondents were willing to answer the questionnaire and fill in all the survey materials including the Main Task Booklet. They found the tasks interesting and complained only about the length of the Main Task Booklet. In 3% of the cases (out of the total 1953) respondents did not fill in the Main Task Booklet or only answered 1-2 questions in the first block. There were further 4% where the Booklet was partially filled in.

Non-response

The issued sample of SSAL2009 covered the whole statistical universe (15-65 years old Scottish residents) without any systematic restrictions of geographic or other nature (beyond the usual preventive conditions of otherwise eligible respondents, such as permanent illness, disability or language). However, due to differential levels of non-response, some segments of the society were under- or overrepresented by the achieved SSAL2009 sample.

Such deviations from the statistical distributions of the sampled universe are part of the reality of any survey research. It is not possible to exactly define the bias that is introduced by deviations (e.g. the extent to which survey results differ from hypothetical results in absence of non-response) that an imperfect response rate and differential non-response patterns introduce in the (unweighted) SSAL2009 results as survey researchers have no access to the vast majority of those respondents who could not be interviewed within the main study to clarify their result and compare it to those interviewed.

Therefore surveys accept that the results the obtain may be considered as solid if a) the overall response rate is satisfactory (SSAL2009 performed acceptably in this regard with a response rate of 59%), and b) if the level of non-response is similar in various key social strata (regions, age groups, etc.). The section below provides detail on the second point, and concludes that the non-response patterns did not show significant irregularities.

Table 6. Non-response patterns, individual level

	SSAL2009 (unweighted)	Total population
SEX	%	%
male	37,8	49,1
female	62,2	50,9
	100	100
AGE		
16-29	27,3	27,6
30-49	41,4	43,5
50-65	31,3	28,9
	100	100
EDUCATION LEVEL		
SVQ[4] 4-5	26,4	33,1
SVQ2-3	53,3	45,9
less/other	20,3	21
	100	100

The *pattern* of non-response in SSAL2009 may be assessed *directly*, along known parameters of each sampled address (region, SES[5], urbanization level) or *indirectly* for parameters which were not known for the sampled addresses but data on the statistical universe exist and can be compared to the properties of the achieved sample.

Starting with the indirect comparisons, the gender, age and education level distributions in the SSAL2009 unweighted sample and the Scottish 15-65 years old resident population are compared (see Table 6.). The results indicate that males are

[4] Scottish Vocational Qualifications

[5] Socio Economic Status

under- and females are overrepresented in the sample. Also, to a lesser extent, the best educated (with SVQ4-5 level completed education) were underrepresented, and those in the SVQ2-3 category were overrepresented in the survey sample, compared to their real proportion in the general population. Deviations remained insignificant in terms of age.

When considering directly comparable characteristics of productive and non-productive addresses in the issued sample, we can also look at the type of non-response, focusing on two distinct categories of refusals (when contact was made but the respondent – or a proxy person in the sampled address – refused to participate) and non-contacts (when the study failed to make contact with the address within the three contact attempts prescribed for each sampled household).

Table 7. Non-response patterns, address level: REGION

NUTS2[6]	Successful interviews %	Refusal %	Non-contact %	*Total population %*
North Eastern Scotland	7.7	7.5	14.3	*10.1*
Eastern Scotland	41.4	36.9	34.5	*38.3*
South Western Scotland	44.8	48.6	41.0	*44.5*
Highlands and Islands	6.1	7.0	10.2	*7.1*
	100	100	100	*100*

Refusals as well as non-contacts were relatively less frequent in Eastern Scotland making it somewhat overrepresented in the achieved sample. On the other hand, an above average non-contact rate in the North East (and to a lesser extent in the Highlands & Islands) resulted in a slight under-representation in these regions. Explicit refusals hindered access to respondents in the South West the most. Generally, the achieved sample did not represent a significant bias in terms of coverage (the largest differences were about 3 percentage points).

Table 8. Non-response patterns, address level: SES

SOCIO-ECONOMIC STATUS (SES)	Successful interviews %	Refusal %	Non-contact %	*Total population %*
first quintile	21.7	16.8	15.1	*19.3*
second quintile	21.4	25.7	19.9	*19.6*
third quintile	16.9	16.7	17.3	*19.8*
fourth quintile	19.2	19.4	18.8	*20.6*
fifth quintile	20.9	21.5	28.9	*20.7*
	100	100	100	*100*

[6] http://www.statistics.gov.uk/geography/nuts_scot.asp

Considering neighbourhoods' socio-economic classification, the achieved sample did not show any significant differences compared to the statistical properties of the universe, however the middle quintile was slightly under-represented.

Table 9. Non-response patterns, address level: Urbanisation Level

URBANISATION	Successful interviews %	Refusal %	Non-contact %	*Total population %*
1 - urban core	40.9	37.2	39.5	*40.1*
2 - urban other	43.7	48.2	41.2	*42.1*
3 - rural	15.4	14.6	19.3	*17.9*
	100	100	100	*100*

SSAL2009 represents the statistical population in terms of urbanisation level fairly well, too, with rural residents slightly underrepresented in the survey sample compared to their real proportion. In this segment non-contact was a more frequent source of non-response, while those residing in Scottish suburbia ('urban other') were more likely than others to refuse participation.

None of these differences are significant to an extent that they could not be easily corrected by post-stratification weighting of the results. For each case, a weight was assigned in a way that the weighted sample composition reflects the respective resident population of Scotland in each of the above dimensions (see next section).

Weighting

The weighting of the SSAL2009 dataset was a two-fold exercise carried out by Gallup. In the first step, a selection probability weighting was performed, which is necessary especially due to built-in disproportionality in selection resulting from the rule that SSAL2009 can only be administered with one random member of each sampled household. Thus, those living in households with more than one person in the eligible age group, have a lower chance to be sampled. This selection bias was corrected with a weight that is the inverse of each sampled person's selection probability, and this constitutes a design weight for each case in the SSAL2009 datafile.

In a further step, a non-response population weighting was carried out as well. As shown above, despite the best efforts, all social surveys / samples suffer from the effects of non-response. As non-response rates vary by social segments, the sample characteristics will reflect such differences as well (i.e. there are usually fewer males and fewer young people in the achieved samples than in the universe.) In this step, weighting "compensates" for the non-response bias that stems from the field execution process.

The following variables were used in the raking procedure[7] (taking the above described design weight as the input weight):

- Age
- Gender
- (NUTS) regions
- Urbanisation level
- Highest level of completed education
- SES[8] of living area (SIMD, Scottish Index of Multiple Deprivation)

When weighting for education, we realized that 31% of respondents answered in A4 that they did not have any qualification. Our assumption is that the question must have been overly complex and respondents already 'forgot' the first part while they processed the second one. (*A4. Do you have any qualifications from school or college, or connected with work or a government training scheme?*). However, most of those persons who answered 'no' to this question previously answered that they had lengthy educational careers (question A3.)

Using a deliberation to code everyone in the group that indicated being without qualifications, those who went to school for 8-12 years were coded into the SVQ2-3 category and those who had an education lasted longer than 12 years in the SVQ4-5 category. Using this approach, a convincingly similar distribution to the Scottish Labour Force Survey was achieved (which served as the source of population data in this regard).

The table on the next page compares the unweighted and weighted sample distributions to the statistical distributions of the universe, for each of the variable and specific classes used for post-stratification weighting.

[7] By raking, the marginal distributions of the auxiliary variables in the sample are adjusted to conform to the population marginals and not the full joint distribution. Raking has been developed to solve such weighting problems where a full joint distribution of the various weighting classes and levels is unavailable, or when a full cross-classification of the auxiliary variables result is a large number of weighting classes with unstable response rates. This procedure performs iterative proportional fitting in contingency table analysis.

[8] Socio Economic Status

Table 10. Comparison of the unweighted and weighted SSAL2009 sample to the statistical universe

GENDER AND AGE	Total population N	%	Unweighted sample interviews	%	Weighted %
males aged 16-29	474,995	14.0	214	11.0	14.0
males aged 30-49	710,656	21.0	303	15.5	21.0
males aged 50-65	476,479	14.1	221	11.3	14.1
females aged 16-29	459,763	13.6	319	16.3	13.6
females aged 30-49	760,848	22.5	505	25.9	22.5
females aged 50-65	498,895	14.8	391	20.0	14.8
	3381636	100	1953	100	100

Source: Scottish Neighbourhood Statistics, 2007

NUTS2	Total population N	%	Unweighted sample interviews	%	Weighted %
North Eastern Scotland	340,996	10.1	150	7.7	10.1
Eastern Scotland	1,296,193	38.3	808	41.4	38.3
South Western Scotland	1,504,243	44.5	875	44.8	44.5
Highlands and Islands	240,204	7.1	120	6.1	7.1
	3381636	100	1953	100	100

Source: Scottish Neighbourhood Statistics, 2007

URBANISATION	Total population N	%	Unweighted sample interviews	%	Weighted %
1 - urban core	1,354,794	40.1	798	40.9	40.1
2 - urban other	1,422,681	42.1	854	43.7	42.1
3 - rural	604,161	17.9	301	15.4	17.9
	3381636	100	1953	100	100

Source: Scottish Neighbourhood Statistics, 2007

SOCIO-ECONOMIC status	Total population N	%	Unweighted sample interviews	%	Weighted %
first quintile	651,266	19.3	424	21.7	19.3
second quintile	662,478	19.6	417	21.4	19.6
third quintile	670,966	19.8	330	16.9	19.8
fourth quintile	696,954	20.6	374	19.2	20.6
fifth quintile	699,972	20.7	408	20.9	20.7
	3381636	100	1953	100	100

Source: Scottish Neighbourhood Statistics, 2007

EDUCATION LEVEL	Total population N	%	Unweighted sample interviews	%	Weighted %
SVQ4-5	1,125,000	33.1	515	26.4	33.1
SVQ2-3	1,561,200	45.9	1,041	53.3	45.9
less/other	714,100	21.0	397	20.3	21.0
	3400300	100	1953	100	100

Source: LFS, Annual Population Survey, Jan-Dec 2008

Fieldwork summary

Given the available resources and time, the fieldwork component of SSAL2009 was carried out successfully. Though it suffered from some expected and unexpected difficulties, most of these could be tackled and the survey was carried out as planned. The significant delay with the launch of the study resulted in an agreed cut-back in terms of target sample size (as the interviewing process could not be 'speeded up' to compensate for the 5 weeks loss due to the complexity of the tasks in the field).

As it became evident after the first few weeks of data collection, it was not a realistic assumption that the survey – even with the lowered target sample size of 2,000 – could be carried out without giving incentives to the respondents. The absence of rules limiting respondents' time for completion of the Main Task Booklets resulted in very lengthy interviews in some cases especially with older respondents. As a result of the longer than expected second part of the interview, sometimes interviewers had difficulties with handling situations when respondents lost their patience at one point in the Main Task Booklet and rushed through the remaining tasks or, on the contrary, respondents spent more than an hour on filling in the booklet.

In most of the surveys interviewers have continuous interaction with respondents during the interview, whilst in the case of SSAL2009, they were supposed to have a more passive role and to just administer the attempts during the self-completion part of the interview (i.e. the test itself). With more family members being present in the household during the interview, the situation could be even more awkward as 3-4 people were sitting there for an hour and all waiting for the respondent to finish the interview.

Only few respondents were aware from the media that the study was taking place. A wider media support would have helped to increase respondents' willingness to take part in the survey.

It proved to be difficult for some of the interviewers to handle all the paper work while administering the visits, keeping track on the booklet distribution (via rotation sheet), copying the right IDs and using the stickers on each and every survey document.

All these issues led to a serious interviewer retention problem preventing the fieldwork from gaining the expected pace.

Below we detail some specific problems that posed barriers to fieldwork execution.

Length of the interview

The interview length was indeed a significant issue. During the field visit, Gallup witnessed that potential respondents' first question when deciding on participation was about the duration of the interview.

If interviewers were honest with the length (saying that it takes more than an hour), respondents often refused to participate. If interviewers were less transparent on the

expected length of the interview and the test, it caused frustration later on if respondent was not even mid-way through the Core Tasks after 30 minutes.

As interviewers did not want to hold up respondents longer than necessary, they usually put the stickers, IDs on each document after leaving the household, the same happened with administering the outcome of visits. This occasionally resulted in a mix-up in the administration of the visits and proper identification of the various materials. Quality control has also found that some of the interviewers began to rush the main questionnaire so as to try and complete the interview within an acceptable time; this in turn created far more human errors within the questionnaire which then increased the amount that had to be declined or refused.

There were two interviewers who attempted to deal with the issue of duration by leaving the materials with the respondents for later collection. This was a very significant breach of protocol. The interviewers were removed from the project and their data destroyed.

The extended duration created a further problem: interviewers were getting paid per interview, so the lengthened process meant that their foreseen compensation was not adequate, and had to be renegotiated after the fieldwork had started.

Sporadic safety concerns

In areas of deprivation it became very difficult to complete more than one interview per evening and many interviewers chose to be accompanied by a 2nd interviewer or member of family to ensure their safety.

Introduction of respondent incentives

Due to the length issues, low response rate and the interviewer retention problem, the fieldwork provider decided to introduce a cash incentive for respondents of £10 for completion of the Main Task Booklets. The incentive was offered only towards the end of the interview as “completion compensation.” Compensation was offered when respondents are about to finish filling in the Main Task Booklet to make both the interviewer and respondent feel better about consuming up to (in extreme cases) 90 minutes of the respondent’s time. This way it could increase the amount of data available for analysis without compromising the sampling protocol.

Although this was not included in the budget, a few weeks of very slow progress and frustration of interviewers led to the realisation that this was needed to help the fieldwork go smoother and decrease soft-refusals in the middle of completing the Main Booklet.

Compensation was not launched at once by all the interviewers but it was introduced gradually:

- March 10th-25th: respondents did not get any compensation
- March 25th-May 8: interviewers started to compensate respondents

- May 8: all respondents got compensation

Administrative Irregularities

During the data processing phase a few cases have been discovered when interviewers recorded 7 or 8 interviews as completing per day. Gallup investigated these cases to verify if these anomalies represent a threat to the validity of the result, if they should be kept in the datafile or should be dismissed.

Checking with the fieldwork agency and the interviewers, we have found that the problems were due to some irregularities in the administration of the survey documents. Interviewers did not always record the date and start-end time of the interview on the interviewing day but made retrospective assumptions as to the approximate dates and times of the interview. In a few cases (per PSU) interviews were broken off before the interviewer could finish the interview, therefore they scheduled these 'partially completed' interviews to be finished on the same day. The interviews did take place in respondents' homes and there were no irregularity in the interviewing process other than the inaccuracy in the time and date recordings.

Gallup's conclusion was that the errors discovered happened as inaccuracies in recording the times and dates rather than as irregularities in the interviews themselves. In addition, it involved only a small number of cases consequently they do not represent a threat to the validity of the results.

ETS ran an analysis of the differences in scores between interviews based on the recorded rate per day. This analysis did not produce any significant differences, with the same distribution of scores showing up consistently. The research team concluded that there was no evidence that the administrative irregularities had affected data quality.

4. SCORING, SCALING, AND STATISTICAL MODELS FOR PROFICIENCY ESTIMATION OF SSAL2009 DATA

Kentaro Yamamoto
Educational Testing Service

The results from the SSAL2009 are reported on three scales upon which the International Adult Literacy Survey (IALS) were reported in 1994-1998. Along with the measurement construct of three literacy scales, the cognitive booklets and data analysis method were also modelled after those used for the IALS (Yamamoto, 1998, Yamamoto & Kirsch, 1998). With scaling methods, the performance of a sample of respondents can be summarized on a series of scales even when different respondents have been administered different assessment items. The two major goals of the SSAL2009 were to report results of adult population in Scotland on the common scales, and to have those scales corresponding to the IALS scales. This section describes the models and procedures used to conduct statistical analyses.

Overview of Analysis

The SSAL2009 gathered descriptive and proficiency information on 1953 sampled respondents through a background questionnaire and a series of assessment booklets containing prose, document, and quantitative literacy tasks. Design and administration of data collection activities were carried out by an agency based upon the description of the study in the various documents supplied from ETS. All survey respondents aged between 16 and 65 who were living in households were potential respondents, i.e. the institutional population was excluded.

Each survey participant answered a common set of background questions concerning his or her demographic characteristics, educational experiences, labour market experiences, and literacy related activities. Responses to these background questions serve two major purposes. First, they provide a way to summarize the survey results using an array of descriptive variables, such as sex, age, educational attainment, and country of birth. Second, they increase the accuracy of the proficiency estimates for various subpopulations, as described later. Background information was designed to be collected by interviewers and the process did not require respondents to read any materials, and thus was independent of the respondents' proficiency. A subset of questions was adapted for current Scotland population from the questions used in 1996.

The remainder of respondents' time was spent completing a booklet of literacy tasks measuring their prose, document, and quantitative skills. The assessment tasks administered in the SSAL2009 were based on the UK adaptation of IALS tasks. All cognitive tasks included in the assessment were open-ended or constructed-responses; these required respondents to provide a written answer.

To achieve good content coverage of each of three literacy domains, the number of tasks in the assessment had to be quite large. Altogether, 114 cognitive tasks were

administered. Yet the time burden for each respondent also needed to be within an acceptable range. To accommodate these two conflicting requirements — in other words, to reduce respondents' time burden without sacrificing good representation of the content domain — each respondent was administered only a fraction of the pool of tasks, using a variant of matrix sampling.

Respondents' literacy proficiencies are estimated based on their performance on the cognitive tasks administered in the assessment. Unlike multiple-choice questions, which are commonly used in large-scale surveys and which offer a fixed number of answer choices, open-ended items such as those used in the SSAL2009 elicit a large variety of responses. Because raw data is seldom useful by itself, responses must be grouped in some way in order to summarize the performance results. As they were scored, and in the ensuing analyses, responses to the SSAL2009 open-ended items were classified into four categories: correct, incorrect, omitted, and not presented.

As noted earlier, a variant of matrix sampling was used in this survey, so that different respondents received different sets of items. Accordingly, in reporting the survey results it is inappropriate to use any statistic based on the number of correct responses, such as the proportion of items answered correctly. Differences in total scores (or statistics based on them) between respondents who took different set of items may be caused by differences in respondents' abilities, differences in difficulty between the two sets of items, or both. Unless one makes very strong assumptions — for example, that the two sets of items are perfectly parallel — the performance of the two groups assessed in a matrix sampling arrangement thus cannot be directly compared using total score statistics. Moreover, item-by-item reporting ignores the similarities of subgroup comparisons that are common across items. Finally, using the average percentage of items answered correctly to estimate the proficiency means of examinees in a given subpopulation does not provide any other information about the distribution of skills within that subpopulation.

These limitations of conventional scoring methods can be overcome by using item response theory (IRT) scaling. When several items require similar skills, the response patterns should have some regularity. This regularity can be used to characterize both respondents and items in terms of a common standard scale, even when all respondents do not take identical sets of items. In this way, it becomes possible to discuss distributions of performance in a population or subpopulation and to estimate the relationships between proficiency and background variables.

It is important to point out that regardless of what procedure is used to aggregate data, a certain amount of information is regarded as non-essential to the analysis. The methods and procedures employed to analyze the SSAL2009 results were carefully designed to capture most of the dominant characteristics of the data.

Scaling and analyses of the SSAL2009 were carried out separately for three domains of literacy: prose, document, and quantitative. By creating a separate scale for each of these areas, it is possible to explore potential differences in subpopulation performance across these areas of literacy.

Scoring

After Gallup and Blauw completed data collection, the instruments were sent to the National Foundation for Educational Research in Slough, where safety copies of all materials were made. Then the entire batch of materials was shipped from the UK to New Jersey for marking and processing by ETS.

The handwritten responses were individually scored. While there had been some concerns that US-based scorers might not always appreciate the meaning of responses collected from UK-based subjects there were no such cases reported. This section addresses scoring practices and issues that arose.

Rescoring reliability

Five scorers with higher than college degree educational credentials were hired and received 3 days of training and practice on scoring. Rescoring was also used to improve quality control of data. Beyond intensive training on scoring of responses to open-ended cognitive items using the scoring manual, it still remains an opportunity to improve accuracy of scoring by aligning scoring criteria through studying unusual responses among scorers. In addition, some scorers may require additional training to become as accurate in scoring as other scorers. Early identification of such scorers is crucial to attain accurate scores overall.

A procedure was set up to monitor scoring accuracy by following a schedule of variable sampling ratio following table 11. First all booklets were bundled together, each bundle including 7 each of 7 booklets. Bundle 39 included 38 booklets. At the beginning of scoring activities, almost all responses were rescored to identify inaccurate scorer and unique or difficult responses that were not in the scoring manual. Average agreements were calculated as average of agreement proportion across all items and monitored frequently. Some precautions were made to ensure independence of the first and the second scores. For example, scorers must be different persons, and the second scorer should not be able to see the scores given by the first scorer.

Since the rescoring was used as a tool to improve quality control, update of rescoring was not made to the database. Thus the following agreements indicate the minimum agreement of the scoring procedures. Sometimes interpretation of the scoring guide was found to be erroneous due to ambiguity of description and required further discussion. If instructions in the scoring manual were found to be ambiguous, changes were made and the first scores reflected such changes but not the second scores. However, comparison with the second scores with remaining errors would underestimate the rescore reliability somewhat.

The amount of the underestimation of the reliability must be very small considering the average reliability was 94%. This was two percentage points lower than the average of all IALS countries rescore reliabilities. Standard classical test analyses were carried out and summarized at block level in table 12.

Table 11: Scoring Design

	Scorers				
Bundle	1	2	3	4	5
1	X	X			
2	X	X			
3	x		x		
4	x		x		
5		x	x		
6		x	x		
7		x		x	
8		x		x	
9			x	x	
10			x	x	
11			x		x
12			x		x
13				X	X
14				X	X
15	X			X	
16	X			X	
17	X				X
18	X				X
19		X			X
20		X			X
21	X				
22	X				
23	X				
24	X				
25		X			
26		X			
27		X			
28		X			
29			X		
30			X		
31			X		
32			X		
33				X	
34				X	
35				X	
36				X	
37					X
38					X
39					X

Table 12: Block Level Average Statistics: SSAL2009

Block #	Average Number Correct	Average Proportion Correct	Average R-Biserial	Average Reliability	Average Proportion Omit	Average Proportion notreached
1	10.51	.72	.55	.82	.04	.02
2	10.03	.70	.46	.81	.05	.04
3	9.31	.65	.53	.83	.05	.05
4	12.88	.74	.50	.83	.04	.03
5	10.63	.66	.57	.86	.05	.06
6	9.69	.77	.54	.81	.03	.03
7	10.26	.71	.55	.83	.05	.04

Evaluation of scoring comparability of SSAL2009 to the international literacy scales was carried out through IRT scaling based on the idea that deviation from scoring consistent to IALS will produce misfit of international common item parameters. Good fit to the international IRT parameters would ensure the inferences based on the scales scores would be comparable to the previous IALS reports.

Scaling Methodology

This section reviews the scaling model employed in the analyses of the SSAL2009 data and explains the multiple imputation or "plausible values" methodology.

The scaling model

The scaling model used for the SSAL2009 is the two-parameter logistic (2PL) model from item response theory (Birnbaum, 1968; Lord, 1980). It is a mathematical model for the probability that a particular person will respond correctly to a particular item from a single domain of items. This probability is given as a function of a parameter characterizing the proficiency of that person, and two parameters characterizing the properties of that item. The following 2PL IRT model was employed in the SSAL2009 (identical to IALS):

$$P(x_{ij} = 1 \mid \theta_j, a_i, b_i) = \frac{1}{1.0 + \exp(-Da_i(\theta_j - b_i))}$$

where:

x_{ij} is the response of person j to item i, 1 if correct and 0 if incorrect;

q_j is the proficiency of person j (note that a person with higher proficiency has a greater probability of responding correctly);

a_i is the slope parameter of item i, characterizing its sensitivity to proficiency;

b_i is its locator parameter, characterizing its difficulty.

Note that this is a monotone increasing function with respect to q; that is, the conditional probability of a correct response increases as the value of q increases. In addition, a linear indeterminacy exists with respect to the values of q_j, a_i, and b_i for a scale defined under the two-parameter model. In other words, for an arbitrary linear transformation of q say $q^* = Mq + X$, the corresponding transformations $a^*_i = a_i/M$ and $b^*_i = Mb_i + X$ give:

$$P(x_{ij} = 1 | \theta^*_j, a^*_i, b^*_i) = P(x_{ij} = 1 | \theta_j, a_i, b_i)$$

To link the IALS scales to the US National Adult Literacy Survey (NALS) scales, the original scale set, the same linear transformation of the scales was used. The scale indeterminacy described above was resolved by setting an origin and unit size of q to the reported scale means and standard deviations of the NALS young adult literacy assessment. The main assumption of IRT is conditional independence. In other words, item response probabilities depend only on q(a measure of proficiency) and the specified item parameters, and not on any demographic characteristics of examinees, or on any other items presented together in a test, or on the survey administration conditions. This enables us to formulate the following joint probability of a particular response pattern x across a set of n items.

$$P(\underline{x} | \theta, \underline{a}, \underline{b}) = \prod_{i=1}^{n} P_i(\theta)^{x_i} (1 - P_i(\theta))^{1-x_i}$$

Replacing the hypothetical response pattern with the real scored data, the above function can be viewed as a likelihood function that is to be maximized with a given set of item parameters. These item parameters were treated as known for the subsequent analyses.

Another assumption of the model is unidimensionality — that is, performance on a set of items is accounted for by a single unidimensional variable. Although this assumption may be too strong, the use of the model is motivated by the need to summarize overall performance parsimoniously within a single domain. Hence, item parameters were estimated for each scale separately.

Testing the assumptions of the IRT model, especially the assumption of conditional independence, is a critical part of the data analyses. The conditional independence means that respondents with the identical ability have a similar probability of producing a correct response on an item regardless of their country membership. This assumption applies to those subsamples in a country who received different set of items. Serious violation of the conditional independence assumption would undermine the accuracy and integrity of the results. It is a common practice to expect a portion of items to be found not suitable for a particular subpopulation. Thus, while the item parameters were being estimated, empirical conditional percentages correct were monitored across the samples. IN the IALS IRT model the percentages of

correct responses obtained by more than three country samples were quite different from the majority of countries for about 10% of items, and these items were dropped from the IALS analyses. For SSAL2009 data analysis the same items were also dropped from the analysis.

Item parameter evaluation

One of the strengths of IRT models is that when their assumptions hold and estimates of the model's item parameters are available for the collections of items that make up the different test forms, all results can be reported directly in terms of the IRT proficiency. This property of IRT scaling removes the need to establish the comparability of number-correct score scales for different forms of the test.

In SSAL2009, the 2PL item-parameters of IALS for each scale were evaluated using a current version of Yamamoto's (1989) Hybil program. Hybil procedures are based on an extension of the marginal-maximum-likelihood approach described by Bock and Aitkin (1981).

$$L(\beta) = \prod_g \prod_{j,g} \int_\theta P(\chi_{j,g} \mid \theta, \beta) f_g(\theta) d(\theta)$$
$$\approx \prod_g \prod_{j,g} \sum_k P(\chi_{j,g} \mid X_k, \beta) A_g(X_k)$$

In the equation, $P(\underline{x}_{j,g}|qb)$ is the conditional probability of observing a response vector $\underline{x}_{jg}$ of person j from a survey g, given proficiency qand vector of item parameters b = $(a_1,b_1,...a_i,b_i)$ and $f_g(q)$ is a population density for q in a survey *g*. The proficiency densities of each survey population were estimated concurrently with item parameters.

The $f_g(q)$ in the above equation are approximated by multinomial distributions over a finite number of "quadrature" points, where X_k, for k=1,..,q, denotes the set of points and $A_g(X_k)$ are the multinomial probabilities at the corresponding points that approximate $f_g(q)$ at q=X_k.

Maximization of L(ß) is carried out by an application of an Expectation-Maximisation algorithm (Dempster, Laird, & Rubin, 1977). When population densities are assumed, known, and held constant during estimation, the algorithm proceeds as follows: In the E-step, provisional estimates of item parameters and the assumed multinomial probabilities are used to estimate "expected sample sizes," at each quadrature point for each group, $\hat{N}_{g,k}$. These same provisional estimates are also used to estimate an "expected frequency" of correct responses at each quadrature point for each group, $\hat{r}_{g,k}$. In the M-step, improved estimates of the item parameters are obtained by treating the $\hat{N}_{g,k}$ and $\hat{r}_{g,k}$ as known and carrying out maximum-likelihood logistic regression analysis to estimate the item parameters ß, subject to any constraints associated with prior distributions specified for ß.

Standardized sample weights were used during item parameter estimation, i.e., sum of weights are constrained to the sample size within every separate surveys. Different subpopulation distributions would be observed within different assessment due to target sampling population and sampling design. By applying post-stratified weights, vital characteristics of the sample can be closely matched to the characteristics of the population.

All cognitive items were estimated jointly on the aggregated data of all survey samples while maintaining the unique distribution of each survey sample. It is known that the samples for each assessment came from somewhat different populations with different characteristics. The calibration procedure should take into account the possibility of systematic interaction of samples and items to estimate unbiased estimates of sample distributions and item parameters. For that reason, a normal distribution with a unique mean and variance for the population of each survey samples was estimated concurrently with item parameters.

There are two options for accommodating the misfit of the IRT model while keeping the common scale intact. One approach is to drop the deviant items from the analysis. A drawback of this option is that it results in a smaller number of items, especially if items are dropped when the IRT functions differ in one or two surveys. We would use this approach if the IRT model did not fit at all, for example, if the response function was negative, or if all observed response functions were so far apart from each other that one set of item parameters would not describe responses from other survey. The approach used in this study was to psychometrically model large deviations by estimating best fitting item parameters.

The common item parameters must fit well in order to justify the use of the item parameter estimates without modification. A graphical method as well as c^2 statistics and square Root of weighted Mean Squared Deviation (RMSD), and weighted Mean Deviation (MD) were used to verify such fit at an item level for every survey separately. Deviations are based on the difference between model-based expected proportions correct and observed proportions correct at each equally spaced 41 ability scale values.

For 34 Prose items, average RMSD was 0.0476 and average MD was 0.0003. For 34 Document items, average RMSD was 0.0387 and average MD was 0.0002. For 33 Quantitative items, average RMSD was 0.0392 and average MD was 0.0010. Number of items that required item parameters different from IALS common parameters was 5 for Prose, 2 for Document and 1 for Quantitative. Table 2 below presents the item parameters for all 101 items. Item parameters unique to SSAL2009 data were italicized in the table. These fits were remarkably good in any conventional standard.

Table 13: Item Parameters Used for SSAL2009

PROSE			DOCUMENT			QUANTITATIVE		
NAME	A	B	NAME	A	B	NAME	A	B
COREQ1S1	0.7192	-2.6678	COREQ2S1	0.5846	-2.4513	COREQ3S1	0.7185	-1.4168
B1Q5S1	0.7799	-0.1044	B1Q1S1	0.5857	-0.3740	COREQ5S1	0.6899	-2.0112
B1Q6S1	0.7738	-0.5154	B1Q2S1	0.7713	-0.7400	B1Q4S1	0.7701	-0.8377
B1Q10S1	1.1668	-1.1003	B1Q13S1	0.5931	-1.3897	B1Q7S1	1.0552	-0.3494
B1Q11S1	0.7019	0.9219	*B2Q8S1*	*0.6367*	*0.8830*	B1Q9S1	0.9228	-0.4251
B2Q1S1	1.1433	-1.0110	B2Q10S1	0.7302	0.1457	B1Q14S1	0.9407	-0.5767
B2Q3S1	1.3148	-1.3733	B2Q111S1	1.4369	-0.6939	B1Q15S1	0.8251	-1.2102
B2Q6S1	0.7665	0.0958	B2Q112S1	1.0205	0.0152	B2Q4S1	0.6666	-0.5153
B2Q7S1	0.8575	1.0713	B2Q113S1	1.4037	-0.7766	B2Q5S1	0.6784	1.2012
B3Q7S1	1.0351	-0.0734	B2Q114S1	2.4093	-0.6466	B2Q9S1	0.9995	-1.2197
B3Q8S1	0.7346	0.2182	*B2Q115S1*	*0.6865*	*0.0561*	B3Q1S1	1.1534	-0.7184
B3Q9S1	1.2476	-0.3169	B3Q2S1	0.7329	0.8619	B3Q3S1	1.2560	-0.6442
B3Q11S1	1.0037	-0.5762	B3Q5S1	0.7878	-0.6396	B3Q6S1	0.9313	-0.9526
B3Q12S1	0.6241	-0.3542	B4Q4S1	1.0883	0.8493	B3Q14S1	1.0197	0.1337
B3Q13S1	0.7329	-0.5843	B4Q5_1S1	1.3161	0.4609	B4Q3S1	0.7627	-1.1543
B3Q15S1	*0.5532*	*-1.1040*	B4Q121S1	0.7467	-1.2600	B4Q5_2S1	1.1412	-0.2589
B4Q1S1	1.0378	-2.3564	B4Q122S1	0.7904	-0.6811	*B4Q9S1*	*0.9677*	*-0.5220*
B4Q2S1	*0.7080*	*0.2953*	B4Q123S1	0.8511	-1.2453	B4Q10S1	1.3771	1.3226
B4Q6S1	0.7136	-0.2751	B4Q124S1	0.9469	-1.6756	B4Q11S1	0.7852	-0.9822
B4Q7S1	*0.8012*	*0.4303*	B5Q7S1	1.0340	-0.6991	B4Q125S1	0.7794	-1.9280
B5Q1S1	*0.7505*	*-2.5736*	B5Q8S1	0.8414	0.0519	B4Q126S1	0.8738	-1.8929
B5Q2S1	0.8922	-1.7507	B5Q10S1	0.9783	0.2543	B5Q9S1	0.7482	-0.8002
B5Q3S1	0.7285	-1.3898	B5Q114S1	0.9103	0.3403	B5Q111S1	0.9347	0.2062
B5Q4S1	0.9761	0.7328	B5Q12S1	0.8047	0.4316	B5Q112S1	1.1306	0.2767
B5Q5S1	0.7591	-0.0229	B6Q4S1	0.9860	-1.1912	B5Q13S1	1.1349	0.3535
B5Q6S1	1.1306	0.1867	B6Q6S1	0.7877	-0.1162	B5Q14S1	1.1114	-0.1646
B6Q1S1	0.9650	-2.0106	B6Q9S1	0.9795	-0.2051	B6Q2S1	1.0290	-0.7279
B6Q7S1	1.1319	-0.6179	B6Q11S1	0.7466	0.0355	B6Q3S1	0.8983	-1.3509
B6Q8S1	*1.3327*	*-0.3917*	B7Q1S1	1.0021	-1.7668	B6Q5S1	0.7185	-0.9560
B7Q10S1	1.4169	-0.5332	B7Q3S1	0.9467	-0.9203	B6Q10S1	0.8411	0.3338
B7Q11S1	0.9562	0.6939	B7Q4S1	0.8431	-0.3487	B7Q2S1	1.0633	-0.9273
B7Q13S1	1.1502	-1.9411	B7Q7S1	0.7763	0.6504	B7Q5S1	0.9267	-0.1367
B7Q14S1	0.7930	-0.5406	B7Q8S1	1.2369	0.2840	B7Q6S1	0.8326	-0.1623
B7Q15S1	0.8977	-0.7924	B7Q9S1	0.9367	2.3733			

Population Modelling

Most cognitive skills testing are concerned with accurately assessing the performance of individual respondents for the purposes of diagnosis, selection, or placement. Regardless of which measurement model is being used, classical test theory or item response theory, the accuracy of these measurements can be improved—that is, the amount of measurement error can be reduced—by increasing the number of items given to the individual. Thus, achievement tests containing more than 70 items are common. Since the uncertainty associated with each q is negligible, the distribution of q or the joint distribution of q with other variables can be approximated using individual qs.

When analyzing the distribution of proficiencies in a group of persons, however, more efficient estimates can be obtained from a sampling design similar to that was used in the IALS. The survey solicits relatively few responses from each sampled respondent while maintaining a wide range of content representation when responses are summed for all respondents. The advantage of estimating population characteristics more efficiently is offset by the inability to make precise statements about individuals. Uncertainty associated with individual $q$estimates is too large to be ignored. Point estimates of proficiency that are, in some sense, optimal for each sampled respondent could lead to seriously biased estimates of population characteristics (Wingersky, Kaplan, & Beaton, 1987.)

Plausible value methodology was developed as a way to estimate key population features consistently and to approximate others no worse than standard IRT procedures would. A detailed review of plausible value methodology is given in Mislevy (1991). Along with theoretical justifications, Mislevy presents comparisons with standard procedures, discusses biases that arise in some secondary analyses, and offers numerical examples. The following is a brief survey of the plausible values approach, focusing on its implementation in the SSAL2009 analyses.

Let y represent the responses of all samples respondents to background questions and questions on engagement to literacy activities, and let q represent the scale proficiency values. If q were known for all sampled examinees, it would be possible to compute a statistic t(q,y)—such as a scale or composite subpopulation sample mean, a sample percentile point, or a sample regression coefficient—to estimate a corresponding population quantity T.

Because the scaling models are latent variable models, however, q values are not observed even for sampled respondents. To overcome this problem, we follow Rubin (1987) by considering q as "missing data" and approximate t(qy) by its expectation given (x,y), the data that actually were observed, as follows:

$$t^*(\underline{x},\underline{y}) = E\left[t(\underline{\theta},\underline{y})\middle|\underline{x},\underline{y}\right]$$
$$= \int t\left(\underline{\theta},\underline{y}\right)p\left(\underline{\theta}\middle|\underline{x},\underline{y}\right)d\theta$$

It is possible to approximate t^* using random draws from the conditional distribution of the scale proficiencies given the item responses x_j, background variables y_j, and model parameters for sampled respondent *j*. These values are referred to as imputations in the sampling literature, and as plausible values in many of population surveys. The value of *q* for any respondent that would enter into the computation of *t* is thus replaced by a randomly selected value from his or her conditional distribution. Rubin (1987) proposed to repeat this process several times so that the uncertainly associated with imputation can be quantified by "multiple imputation." For example, the average of multiple estimates of *t*, each computed from a different set of plausible values, is a numerical approximation of t^* of the above equation; the variance among them reflects uncertainly due to not observing *q*. It should be noted that this variance does not include the variability of sampling from the population.

It cannot be emphasized too strongly that plausible values are not test scores for individuals in the usual sense. Plausible values are only intermediary computations for calculating integrals of the form of the above equation in order to estimate population characteristics. When the underlying model is correctly specified, plausible values will provide consistent estimates of population characteristics, even though they are not generally unbiased estimates of the proficiencies of the individuals with whom they are associated. The key idea lies in a contrast between plausible values and the more familiar ability estimates of educational measurement that are in some sense optimal for each respondent (e.g., maximum likelihood estimates, which are consistent estimates of a respondent's *q*and Bayes estimates, which provide minimum mean-squared errors with respect to a reference population). Point estimates that are optimal for individual respondents have distributions that can produce decidedly nonoptimal (inconsistent) estimates of population characteristics (Little & Rubin, 1983). Plausible values, on the other hand, are constructed explicitly to provide consistent estimates of population effects. For further discussion, see Mislevy, Beaton, Kaplan, and Sheehan (1992).

Plausible values for each respondent *j* are drawn from the conditional distribution $P(q_j|x_j,y_j,\Gamma,\Sigma)$,where G is a matrix of regression coefficients and Σis a common variance matrix for residuals. Using standard rules of probability, the conditional probability of proficiency can be represented as follows

$$\begin{aligned} P(\underline{\theta}_j \left| x_j, y_j, \Gamma, \Sigma \right.) &\propto P(x_j \left| \underline{\theta}_j, y_j, \Gamma, \Sigma \right.) P(\underline{\theta}_j \left| y_j, \Gamma, \Sigma \right.) \\ &= P(x_j \left| \underline{\theta}_j \right.) P(\underline{\theta}_j \left| y_j, \Gamma, \Sigma \right.) \end{aligned} \qquad (1)$$

where q_j is a vector of three scale values, $P(x_j|q_j)$ is the product over the scales of the independent likelihoods induced by responses to items within each scale, and $P(q_j|y_j,\Gamma,\Sigma)$ is the multivariate joint density of proficiencies of the scales, conditional on the observed value y_j of background responses and parameters Γand S. Item parameters estimates are fixed and regarded as population values in the computation described in this section.

In the analyses of the SSAL2009, a normal multivariate distribution was assumed for $P(q_j|y_j,G,S)$, with a common variance, S, and with a mean given by a linear model with slope parameters, G, based on the first approximately principal components of several hundred selected main effects and two-way interactions of the complete vector of background variables. The background variables embodied included sex, ethnicity, region of the country, respondent education, parental education, occupation, and reading practices, among others. Based on the principal component method, components representing 99 percent of the variance present in the data were selected. The included principal components will be referred to as the conditioning variables, and denoted as y^c. (The complete set of original background variables used in the analyses are listed in the appendices.) The following model was fit to the data.

$$\theta = \Gamma' y^c + \varepsilon$$

where e is normally distributed with mean zero and variance S. As in a regression analysis, G is a matrix each of whose columns is the effects for one scale and S is the three-by-three matrix variance of residuals between scales.

Note that in order to be strictly correct for all functions G of q it is necessary that $p(q|\boldsymbol{y})$ be correctly specified for all background variables in the survey. In the SSAL2009, however, principal component scores based on the nearly all background variables were used. The computation of marginal means and percentile points of q for these variables is nearly optimal. Estimates of functions T involving background variables not conditioned on in this manner are subject to estimation error due to misspecification. The nature of these errors was discussed in detail in Mislevy (1991). Their magnitudes diminish as each respondent provides more cognitive data—that is, responds to a greater number of items. Indications are that the magnitude of these errors is negligible in the SSAL2009 (e.g., biases in regression coefficients below 5 percent) due to the larger numbers of cognitive items presented to each respondent in the survey (on average, 16 items per scale). The exception is the very small sample of respondents who did not proceed beyond the background questions.

These respondents did not attempt the assessment tasks due to their inability to read or write English, physical disability, mental disability, or refusal to participate in the survey. If these respondents had been excluded from the survey, the proficiency scores of some subpopulations in the SSAL2009 would have been severely overestimated and the picture of the nation's literacy skills would have been distorted. These respondents possess few literacy skills, if any, in the language of assessment, and detailed analyses of their proficiencies, not surprisingly, may lead to unstable results.

The basic method for estimating G and S with the EM procedure was described in Mislevy (1985) for a single scale case. The EM algorithm requires the computation of the mean, q and variance, S, of the posterior distribution in equation (1). For the multiple scales of SSAL2009, the computer program DGROUP was used. The program implemented a method to compute the moments using higher order

asymptotic corrections to a normal approximation. Case weights were employed in this step.

After completing the EM algorithm, the plausible values are drawn in a three-step process from the joint distribution of the values of Ģfor all sampled respondents with more than four cognitive items attempted. First, a value of G is drawn from a normal approximation to $P(G\dot{S}|x_j,y_j)$ that fixes S at the value S^ (Thomas, 1993). Second, conditional on the generated value of G (and the fixed value of S=S^), the mean *q*, and variance S_j^p of the posterior distribution in the equation (1) are computed using the same methods applied in the EM algorithm. In the third step, the *q* are drawn independently from a multivariate normal distribution with mean *q* and variance S_j^p . These three steps are repeated five times, producing five imputations of *q* for each sampled respondent.

For those with an insufficient number of responses, the G and Ss described in the previous paragraph were fixed. Hence, all respondents—regardless of the number of items attempted—were assigned a set of plausible values for the three scales. The plausible values could then be employed to evaluate equation 1 for an arbitrary function T as follows:

1. Using the first vector of plausible values for each respondent, evaluate T as if the plausible values were the true values of q. Denote the result T_1.
2. In the same manner as in step 1 above, evaluate the sampling variance of T, or $Var(T_1)$, with respect to respondents' first vectors of plausible values. Denote the result Var_1.
3. Carry out steps 1 and 2 for the second through fifth vectors of plausible values, thus obtaining T_u and Var_u for u=2, . . .,5.
4. The best estimate of T obtainable from the plausible values is the average of the five values obtained from the different sets of plausible values:

$$T_{\cdot} = \frac{\sum T_u}{5}$$

5. An estimate of the variance of T. is the sum of two components: an estimate of $Var(T_u)$ obtained as in step 4 and the variance among the T_us:

$$Var(T_{\cdot}) = \frac{\sum Var_u}{5} + (1 + \frac{1}{5})\frac{\sum (T_u - T_{\cdot})^2}{5 - 1}.$$

The first component in *Var(T.)* reflects uncertainty due to sampling respondents from the population; the second component reflects uncertainty due to the fact that sampled respondents' *q*s are not known precisely, but only indirectly through *x* and *y*.

Linking the IALS Scale to the NALS Scale

The plausible values out of population modelling on the provisional scale must be transformed to the IALS scale for comparison and further analyses. Since SSAL2009 scales were explicitly linked by using the identical item parameters as IALS, the same transformation constants can be applied following:q=Aq*+B where q* is the provisional scale from item calibration and q is the reported scale. Table 14 presents the transformation constants and the mean and standard deviations for the distributions of the three scales.

Table 14: Transformation Constants Applied to Provisional Scale to Produce Reported Scale

Literacy scale	A	B
Prose	51.67	269.16
Document	52.46	237.50
Quantitative	54.41	276.87

Statistical Tests

Analysis of Plausible Values

Plausible values methodology was used in this survey to increase the accuracy of the estimates of the proficiency distributions for various subpopulations and for the adult population as a whole. This method correctly retains the uncertainty associated with proficiency estimates for individual respondents by using multiple imputed proficiency values rather than assuming that this type of uncertainty is zero—a more common practice. Retaining this component of uncertainty requires that additional analysis procedures be used to estimate respondents' proficiencies.

If q values were observed for sampled respondents, the statistic $(t-T)/U^{1/2}$ would follow a t-distribution with d degrees of freedom. Then the incomplete-data statistic $(t^*-T)/(Var(t^*))^{1/2}$ is approximately t-distributed, with degrees of freedom given by

$$v = \frac{1}{\frac{f_M^2}{M-1} + \frac{(1-f_M)^2}{d}}$$

where f_M is the proportion of total variance due to not observing qvalues:

$$f_M = \frac{(1+\frac{1}{M})B_M}{V_M}$$

When B_M is small relative to U^*, the reference distribution for incomplete-data statistics differs little from the reference distribution for the corresponding complete-data statistics. This is the case in the US National Assessment of Educational Progress surveys. If, in addition, d is large, the normal approximation can be used instead of the t-distribution.

For k-dimensional t, such as the k coefficients in a multiple regression analysis, each U_M and U^* is a covariance matrix, and B_M is an average of squares and cross-products rather than simply an average of squares. In this case, the quantity $(T-t^*)V^{-1}(T-t^*)'$ is approximately F distributed with degrees of freedom equal to k and n, with ṇdefined as above but with a matrix generalization of f_M

$$f_M = \frac{(1-M^{-1})Trace(B_M V_M^{-1})}{k}$$

A chi-square distribution with k degrees of freedom can be used in place of f_M for the same reason that the normal distribution can approximate the t distribution.

Statistics t*, the estimates of ability and background variables, are consistent estimates of the corresponding population values T, as long as background variables are included in the conditioning variables. The consequences of violating this restriction are described by Beaton and Johnson (1990), Mislevy (1991), and Mislevy and Sheehan (1987). To avoid such biases, the NALS analysis included nearly all background variables. These variables were orthogonally coded, thus avoiding the necessity of linear coding. This increased the number of variables substantially, however. To capture most of the variances in the background questions with a limited number of variables, principal components were used. Because each subpopulation can have unique relationships among the background variables, one set of principal components is not sufficient for all samples included in the NALS (i.e., the older adult, prison, and household samples). Each set of principal components was selected to include 99 percent of the variance in the background variables. Mislevy (1990) shows that this puts an upper bound of 1 percent on the average bias for all analyses involving the original conditioning variables.

Partitioning the Estimation Error Variance: A Numerical Example

The five plausible values of 3 scales of SSAL2009 can be used to estimate standard errors of measurement only. The SSAL2009 data user need to use some other method to estimate other portion of SE, namely due to sampling of respondents. This section offers an example of the use of multiple plausible values in the NALS analysis to partition the error variance in order to show the relative size of two parts of errors. The proportion of two parts of SE is never uniform and not ignorable. Sometimes, the SE is mainly attributable to sampling errors and sometimes to measurement errors. Ignoring one or the other would results in inaccurate inferences of results.

Table 15 presents data for three subgroups of respondents with differing educational attainments: those whose highest level of education was a GED, a high school

diploma, and a four-year college degree. As noted earlier, five plausible values were calculated for each respondent for each scale. Each column presents the means of these five values.

Table 15: Mean plausible values by level of education

Subsample	N	Five mean plausible values 1	2	3	4	5	Mean of 5 PV's	Sampling error $Var(JK_1)$	Measurement error 1.2* Var(mean)	Total error variance ÖT var.
GED	1062	269.3	268.1	267.9	268.2	267.7	268.2	2.888	0.580	1.86
High school	6107	270.2	270.4	270.3	270.5	270.2	270.3	1.050	0.216	1.13
4 year college	2534	321.2	321.7	322.4	322.8	320.4	321.7	1.408	1.232	1.62

Variance in the mean plausible values is similar but not identical for the three subgroups. As noted previously, variance reflects a component of error attributable to the lack of precision of the measurement instrument and a component of error attributable to sampling. Variance can be reduced by either increasing the precision of the measurement instrument (for example, expanding the number of items) or increasing the size of the sample. The jackknife method was used to estimate error variance due to sampling using the first set of imputed values. This component of variance is expected to be consistent across five imputed values, and the size is influenced by the homogeneity of proficiencies among respondents in a subgroup but not by the sample size or by the precision of the survey instruments. Error variance due to sampling is smaller when the subgroup consists of respondents with similar proficiencies. As noted earlier, total error variance can be calculated as, summation of sampling error variance and 1.2 times measurement error variance.

Despite a relatively large sample size, the mean for respondents with four-year college degrees has a larger error variance than those for the other education groups. In fact, it is twice as large as the variance for respondents whose highest level of education is a General Educational Development (GED) Diploma. The higher variance for this best-educated group is due to the characteristics of the assessment, which encompassed the entire adult population in this country, age 16 and older, and measured a wide range of skills. The precision of the assessment is optimal at the middle of the proficiency range, since that is where most of the population is expected to perform. Since most respondents with four-year college degrees scored above this range, variance due to lack of precision in measurement is quite high. Increasing the sample size would therefore not do much to reduce the variance component for this group. On the other hand, the error variance due to sampling is twice as large for the GED group than for the four-year college degree group indicating their heterogeneity of skills in the group defined by GED.

The last column presents the standard error of the subpopulation mean, which is equal to the square root of the sum of the two components of error variance. The differences among the means can be compared using these standard errors. In doing so, it is first necessary to decide how many comparisons are being made. In this example we might be interested in making three comparisons: GED vs. high school; high school vs. four-year college degree; and GED vs. four-year college degree. Following the Bonferroni method of multiple comparisons, any comparison among these three with a standardized difference greater than 2.39—($mean_1$-$mean_2$)/sqrt(se_1^2+se_2^2), (z_p=0.025/3)—can be considered statistically significant. The difference in means between GED recipients and high-school graduates is not statistically significant (0.97) at the .05 level, but the differences between these two groups and respondents with four-year degrees are significant (22.2 and 26.8 respectively).

Minimum Sample Sizes for Reporting Subgroup Results

The sample sizes were not always large enough to permit accurate estimates of proficiency and/or background results for one or more categories of variables. For results to be reported for any subgroup, a minimum sample size of 60 was required. This number was arrived at by determining the sample size needed to detect an effect size of 0.2 with a probability of 0.8 or greater using a design effect of 2. The design effect 2 implies a sample design-based variance twice that of simple random sampling. The effect size of 0.2 pertains to the true difference in mean proficiency between the subgroup in question and the total population, divided by the standard deviation of proficiency in the total population. An effect size of 0.2 was chosen following Cohen (1988), who classifies effect size of this magnitude as "medium."

Multiple Comparisons Means of Subgroups

The most common comparisons one may want to make using population survey results is to compare proficiency means of various subpopulations, characterized by level of Education, Age, race/ethnicity, region of residence or country of residence among others. The primary comparison is to compare the difference of means of two subgroups. The statistic to evaluate statistical significance of difference of means for two subgroups A and B is t statistics and the formula for that is:

$$t_{df} = \frac{\overline{\theta}_{A,0} - \overline{\theta}_{B,0}}{\sqrt{SE_A^2 + SE_B^2}}$$

Where $\theta_{A,0}$, $\theta_{B,0}$ are the proficiency means of a subgroups A, and B and SE_A, SE_B are the standard errors for the subgroups A, and B including both portions of errors. The degree of freedom for this statistic is defined by a Satterthwaite approximation as follows:

$$df = \frac{\left(SE_A^2 + SE_B^2\right)^2}{\frac{SE_A^4}{df_A} + \frac{SE_B^4}{df_B}}$$

Where df_A is defined as:

$$df_A = \left(3.16 - \frac{2.77}{\sqrt{M}}\right) \frac{\left(\sum_{j=1}^{J} \left(\overline{\theta}_{A,j} - \overline{\theta}_{A,0}\right)^2\right)^2}{\sum_{j=1}^{J} \left(\overline{\theta}_{A,j} - \overline{\theta}_{A,0}\right)^4}$$

Where M is the number of non-empty PSU pairs and it is limited by J, total number of PSU pairs, i.e., the number of replicate weights. Significance levels associated with each pairwise t statistics will be evaluated for their significance for the comparisons of more than two groups.

Whatever the comparison might be, repeating simple pairwise comparisons of multiple groups would find "statistically significant difference" more often than it should. The total number of comparisons (M) of m categories variable is m*(m-1)/2 pairwise comparisons. For example, if p=0.05 is used to define significance level and m=10, we should expect to find 2 out of M=45 comparisons by random chance alone. This incidental finding of significant difference due to large number of pairwise comparisons have been dealt with through adjusting significance level. The most notable procedure is Bonferoni method that uses the target significance level divided by the number of comparisons. This method uses the single adjusted significance level for all comparisons. However, this leads to very stringent criteria and under-identifies the significant differences due to treating all comparisons to be independent.

Hochberg developed alternative method that utilizes the order of significance levels among all comparisons (Hochberg, 1988). The Hotchberg Stagewise Procedure (HSP) starts out with placing the comparisons by increasing order of significance levels, i.e., $P_1 \leq P_2 \leq \ldots \leq P_3 \leq \ldots P_M$. Proceed to sequentially evaluate P_j with adjusted critical significance level of $\alpha/(m-j+1)$ where α is the target significance level, if P_j is smaller than the critical significance level then continue until non-significance comparison is found. All preceding comparisons before the first non-significant comparison are declared significant and all subsequent comparisons will be declared non-significant. Both Bonferoni and HSP both control the type 1 error of false discovery of significant comparison when in fact it is non-significant. The False Discovery Rate (FDR) procedure (Benjamini and Hochberg, 1995) controls the expected proportion of falsely rejected hypotheses, finding the comparison non-significant when in fact it is significant. The procedure is very similar to HSP for ordering the comparisons by the significance level, then use the critical significance level of α*j/m for j-th comparison. Determination of the significance of comparisons is identical to the HSP.

5. THE ANALYSIS OF LITERACIES LEVELS

Dougal Hutchison
National Foundation for Educational Research

As is conventional in educational research, the uncertainty in the results is referred to as error. Note however that the term 'error' as used here is a technical one. It gives an indication of the extent to which a separate parallel administration of the study would have given a different answer. It should not be taken as being equivalent to 'errors' or 'mistakes' on the part of those administering the study. While these are inevitably possible as part of the overall potential variability, the careful quality assurance procedures documented elsewhere in the report of this study mean that it is likely that such a contribution would have a relatively small impact on the over result. The main contributions arise from:

a) Sampling only part of the population of 16-65 adults in the survey
b) Relatively small part of each respondent's literacy assessed.

These are considered in turn.

Sampling contribution to standard error

Sampling only part of the population of 16-65 adults in the survey

The most important component of random error is sampling error, which is the error that arises because the estimate is based on a sample survey rather than a full census of the population. The results obtained for any single sample may, by chance, vary from the true values for the population but the variation would be expected to average to zero over a number of repeats of the survey. The amount of variation depends on the sample size and sample design, the weighting method and the variability of the population on the characteristic of interest. The larger the sample size the lower the sampling error is likely to be.

The standard error (or sampling error) can be estimated directly from the values obtained for the survey sample. This allows calculation of confidence intervals that give an indication of the range in which the true population value is likely to fall. It is usual practice to refer to the 95% confidence interval around a survey value, which is calculated as 1.96 times the standard error on either side of the estimated percentage or mean. The 95% confidence interval for a sample percentage estimate (p) is given by the formula: $p \pm 1.96 * se(p)$

For results based on a simple random sample with no weighting, the calculation of the standard error is straightforward. In the case of a percentage *(p)* for a sub-sample of size *n* the formula is:

$$se(p) = \frac{\sqrt{p(100 - p)}}{n}$$

Where, as here, the sample is stratified the calculation of the standard error is more complex. The Scottish Survey of Adult Literacies, as is currently usual in important large-scale attainment surveys, employed a more complex and administratively efficient design, first taking as primary sampling units a number of postcode areas sampled from the whole country, and then sampling individuals within the selected areas. The sampling procedure is described in more detail in chapter 3, but involved sorting all areas in the country according to relevant characteristics, and then selecting areas at a fixed interval from the sorted list, starting from a randomly selected starting point.

The calculation of sampling errors needs to take account of the complex sample design. Thus, the standard error derived using the above formula should be multiplied by an appropriate design factor *(deft)* which allows for the stratification and clustering in the sample design and also the weighting used in the survey. The design factor is the ratio of the standard error allowing for the complex sample design to the standard error with a simple random sample of the same size.

Recently, the estimation of standard errors by large scale national or international studies has been carried out using the Jackknife method (see e.g. Murray, Kirsch and Jenkins, 1998) or the Balanced Repeated Replications method (OECD, 2005). In this study the approach of earlier Adult Literacy Studies was followed and a Jackknife approach was adopted (see, e.g. Ross, K. 2005). The Jackknife is a computationally intensive procedure, and previous UK implementations have used approximations (Carey, Lowe and Hansbro, 1997; Scottish Executive, 2001): the inexorable increase in computing power has enabled us to perform a closer implementation of the procedure. The jackknifing procedure as implemented in Murray et. al (1998) used the Johnson and Rust (1992) approximation (Yamamoto, 2010). The procedure used here was slightly different.

The primary sampling units (PSUs), here the selected areas, were first ordered, and then paired as in the sampling design. Each pair was regarded as members of a pseudo-stratum for variance estimation purposes. This was done by working through the list of sampled areas in the order in which they were selected and assigning the first and second participating PSUs to the first sampling stratum, the third and fourth PSUs to the second stratum, and so on.

Taking the first stratum as an illustration, the first PSU was assigned a weight of 2, and the second PSU a weight of zero, on top of the weight previously estimated. All PSUs in other strata were given a weight of 1, and the statistic of interest calculated on the modified sample. The process is then repeated, but this time giving the first PSU was assigned a weight of zero, and the second PSU a weight of 2. This entire process is repeated for each of the pseudo-strata, and the variation between the

estimates for each of the jackknife replicate samples and the original sample estimate is used to calculate the jackknife estimate of the sampling error of the statistic.

The formula for the sampling variance estimate of a statistic *t* is given by the following equation:

$$Var_{jrr}(t) = \frac{1}{2}\sum_{h=1}^{H}[t(J_h) - t(S)]^2$$

where *H* is the total number of sampling zones in the sample of the country under consideration. The term *t(S)* corresponds to the statistic of interest for the whole sample computed with the overall sampling weights. The term $t(J_h)$ denotes the same statistic using the h^{th} jackknife replicate sample and its set of replicate sampling weights, which are identical to the overall sampling weights, except for the respondents in the h^{th} sampling zone. For the respondents in the h^{th} zone, all respondents were removed, and the respondents belonging to another PSU were included twice. In practice, this was accomplished by recoding to zero the weights for the respondents to be excluded from the replication, and multiplying by two the weights of an equal number of remaining respondents. Each sampled respondent was assigned a vector of *H* replicate sampling weights W_{hi}. If W_{oi} was the overall sampling weight of respondent *i*, the *h* replicate weights for that respondent were computed as $W_{hi} = W_{oi} * k_{hi}$ where

k_{hi} =2 if the PSU is the 'selected' one in that pseudo-stratum

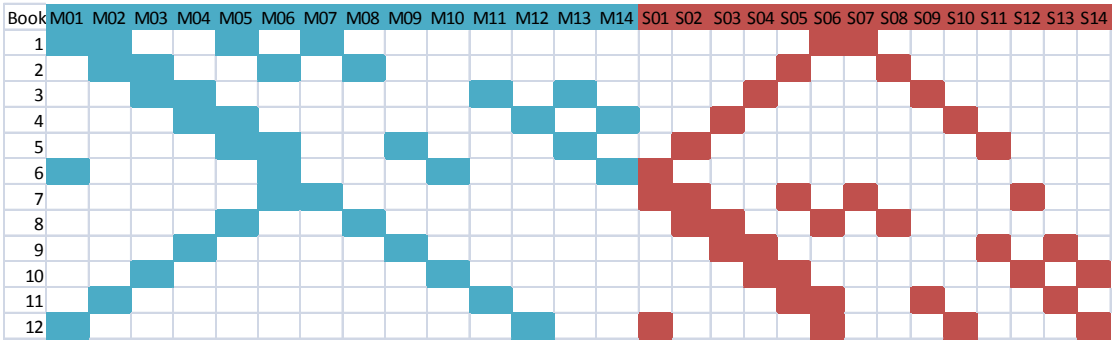
k_{hi} =0 if the PSU is the 'non-selected' one in that pseudo-stratum

k_{hi} =1 otherwise.

Imputation contribution to standard error

Because of the large number of questions involved in the study, it was not feasible to give the entire test to each respondent. Instead a technique known as matrix sampling was used so that all items were answered, but not by any single respondent. Items were arranged in booklets, and each item was 'linked', directly or indirectly, with every other item in the study. Some items were linked directly in that a part of the population took both. Some items were linked indirectly, via other items. An example of a linking arrangement of booklets, for the Trends in International Mathematics and Science Survey (TIMSS), is shown below. This shows two subjects, mathematics and science, each in 14 sub-tests, linked in 12 booklets.

Figure 1: Example of matrix sampling



The details of this structure mean that there is a relatively large degree of random error in the original scores as derived. There is also the possibility of bias arising in the comparisons of groups within the study. For this reason two further steps were taken by ETS in preparing the data. First the scores were conditioned on a very wide range of background information to reduce bias as well as measurement error, and then the remaining measurement error was simulated by producing multiple imputations of the scores. These multiple imputations, five in number, are referred to as plausible values. It is important to note that these plausible values are not equivalent to individual scores, but give superior results for aggregated statistical comparisons. The process is described in more detail elsewhere in this report (Yamamoto, 2009).

Combining sampling and imputation contributions

The best estimate of *T* obtainable from the plausible values is the average of the five values obtained from the plausible values.

$$T = \frac{\Sigma_u T_u}{5}$$

An estimate of the variance of *T* is the sum of two components, an estimate of $Var(T_u)$ and the variance among the T_u s:

$$Var(T) = \frac{\Sigma_u Var_u}{5} + \frac{\left(1+\frac{1}{5}\right)(\Sigma_u (T_u - T)^2)}{5-1}$$

The standard error of *T* is given by

$$SE(T) = sqrt(Var(T))$$

To estimate the design factor, the standard error is calculated under the (counterfactual) assumption that the sample was a simple random sample without plausible values. If this standard error is designated $SE_{srs}(T)$ then the design factor deft is given by the ratio

$$deft = \frac{SE(T)}{SE_{srs}(T)}$$

Application of final sampling error and design factor results

The true standard errors for a selection of the survey estimates are presented in the *SSAL2009 Report of Findings*. For other estimates the 95% confidence interval for a percentage from the Survey can be calculated as:

$$p \pm 1.96 \times deft \times se_{srs}(p)$$

where $se_{srs}(p)$ is the standard error assuming a simple random sample. There may seem to be a degree of circularity here in dividing by the *srs* standard error, and then using it to multiply the same quantity, but this formula can be used for other variable, to avoid the lengthy and complex procedure involved in calculating the full value. An appropriate value of *deft* can be taken from those given in the *Report of Findings* by selecting a variable which is likely to be clustered in the same way. It should be noted that design factors for estimates based on subsamples are generally smaller than those for estimates based on the total sample.

Regression results

This covers the regressions with variables treated as continuous, i.e. the scores on the three literacy scales, prose, document and quantitative, and also the 'odds' estimates. The process involved was essentially similar to that described above for Tables A2 to A7 with some small additional features. For the continuous outcomes, a variance components multilevel model (Goldstein, 2003) was used. For the i^{th} individual in j^{th} primary sampling unit (PSU), the theoretical relationship between the scale score and background characteristics x_{ij} is given by

$$Y_{ij} = \beta_0 + \sum_j \beta_j x_{ij} + u_j + e_{ij}$$

Where u_j is a term random at level 2 (PSU) and e_{ij} is random at individual level. This equation, and more specifically its standard errors, are estimated by carrying out the jackknifing procedure separately for each of the 5 PVs, following the procedure recommended in the Help section of the program MLwiN (Rasbash *et al*, 2009). Since the jackknifing procedure involves reweighting units at PSU level, the jackknife contribution to the weighting was input at level 2. The odds estimates were carried out comparably but using a logistic regression procedure for binary outcomes.

References

Beaton, A. E., & Johnson, E. G. (1990). The average response method of scaling. *Journal of Educational Statistics,* 15, 9-38.

Benjamini, Y.& Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, (57), 289-300.

Birnbaum, A. (1968) Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley Publishing.

Carey, S., Lowe, S., and Hansbro, J. (1997) *Adult Literacy in Britain.* London: The Stationery Office.

Cohen (1988). *Statistical power analysis for the behavioral sciences.* (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum Associates.

Goldstein, H. (2003) *Multilevel models: third edition*. Arnold: London.

Hochberg, Y., 1988 A sharper Bonferroni procedure for multiple tests of significance. *Biometrika,* 75, 800-802.

Johnson, E. G., & Rust, K. F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics, 17* (2), 175-190.

Little, R.J.A. & Rubin, D. B. (1983). On jointly estimating parameters and missing data. *American Statistician,* 37, 218-220.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum Associates.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80,* 993-97.

Mislevy, R.J. (1990). Scaling procedures. In E.G. Johnson and R. Zwick, *Focusing the new design: the NAEP 1988 technical report* (No. 19-TR-20). Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Mislevy, R.J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196.

Mislevy, R.J., Beaton, A., Kaplan, B.A., and Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29* (2), 133-161.

Mislevy, R. J. & Bock, R. D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Morresville, IN: Scientific Software.

Mislevy, R. J. & Sheehan, K. (1987) Marginal estimation procedures. In A. E. Beaton, *Implementing the new design: The NAEP 1983-84 technical report* (pp. 293-360). (no. 15-TR-20) Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Murray, T.S., Kirsch I.S. & Jenkins, L.B. (1998) *Adult Literacy in OECD Countries: Technical Report on the First International Adult Literacy Survey.* Washington, DC: National Centre for Education Statistics.

OECD (2005) *PISA 2003 Data Analysis Manual: SPSS® Users.* Paris: Author.

Rasbash, J., Browne, W., Healy, M., Cameron, B. & Charlton, C. (2009) *MLwiN version 2.15 (computer program).* University of Bristol, Centre for Multilevel Modelling.

Ross, K. (2005) *Quantitative Methods in Research Planning: Module 3: sample design for educational survey research.* Paris: Unesco.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Scottish Executive (2001) *Adult literacy in Scotland: analysis of data from the 1996 Adult Literacy Survey.* Edinburgh: Author.

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computations and Graphical Statistics*, 2, 309-22.

Wingersky, M, Kaplan, B. A., & Beaton, A. E. (1987). Joint estimation procedures. In A. E. Beaton, *Implementing the new design: The NAEP 1983-84 technical report* (pp.285-92) (No. 15-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Yamamoto, K. (1998). Scoring, scaling, and statistical models for proficiency estimation of the IALS. *International Adult Literacy Survey Technical Report.* Ottawa, Canada: Statistics Canada.

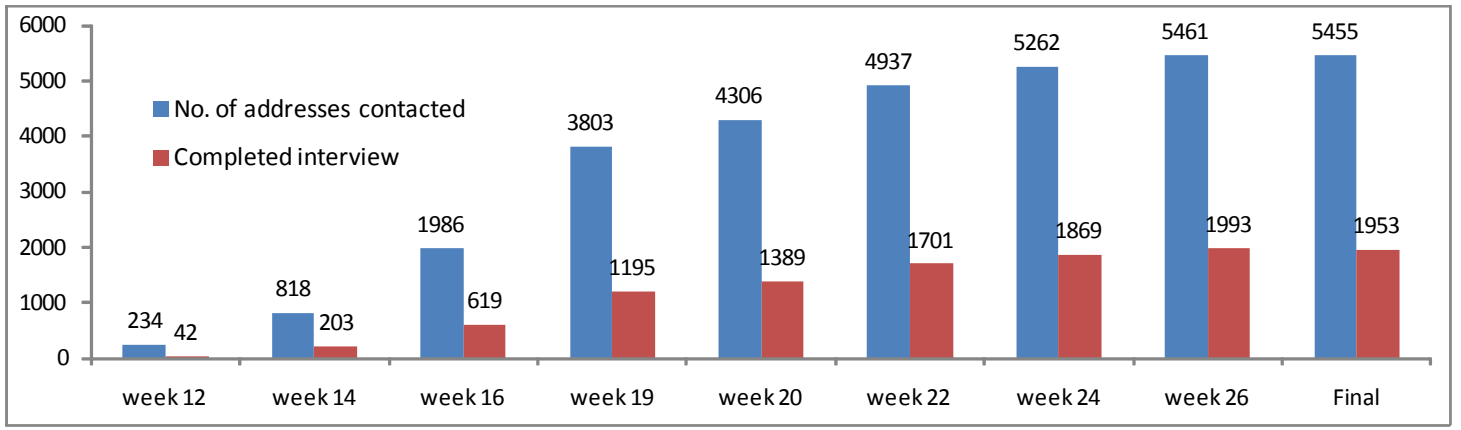
Yamamoto, K & Kirsch, I. (1998). Estimating literacy proficiencies with and without cognitive data of IALS. *International Adult Literacy Survey Technical Report.* Ottawa, Canada: Statistics Canada.

Yamamoto, K., (2010) Personal communication.

APPENDICES

APPENDIX A. RATE OF INTERVIEW COMPLETION DURING FIELDWORK

	week 12	week 14	week 16	week 19	week 20	week 22	week 24	week 26	Final
	20-Mar	03-Apr	17-Apr	05-May	14-May	28-May	10-Jun	24-Jun	1-July
No. of addresses contacted	**234**	**818**	**1986**	**3803**	**4306**	**4937**	**5262**	**5461**	**5455**
Unsuccessful contacts:	**167**	**545**	**1312**	**2544**	**2841**	**3203**	**3368**	**3451**	**3359**
A) Respondent refused	*51*	*168*	422	773	898	1092	1109	1146	1239
B) Respondent was not eligible	*36*	*128*	365	652	738	876	898	940	910
C) Visits made but no contact established (address non-residential, no one at home, etc.)	*80*	249	517	1081	1151	1173	1298	1300	1152
D) Wrong addresses			*8*	*38*	*54*	*62*	*63*	*65*	*58*
Successful contacts:	**67**	**273**	**674**	**1259**	**1465**	**1734**	**1894**	**2010**	**1953**
E) Completed interview	*42*	*203*	619	1195	1389	1701	1869	1993	1953[9]
F) Appointments	*25*	*70*	55	64	76	33	25	17	



[9] Please note that after the final deadline of the fieldwork Blauw delivered additional 58 interviews to Gallup.

B. WEIGHTED AND UNWEIGHTED ITEM PERCENT CORRECT

Percentages of correct responses for each cognitive task question before (UW) and after (W) weighting.

Item ID	Scale	UW	W	Item ID	Scale	UW	W	Item ID	Scale	UW	W
COREQ1S1	P	99	99	B3Q1S1	Q	81	82	B5Q7S1	D	86	88
COREQ2S1	D	98	98	B3Q2S1	D	46	47	B5Q8S1	D	64	64
COREQ3S1	Q	94	94	B3Q3S1	Q	79	80	B5Q9S1	Q	74	75
COREQ4S1	D	97	98	B3Q4S1	D	90	91	B5Q10S1	D	77	78
COREQ5S1	Q	90	90	B3Q5S1	D	86	86	B5Q111S1	Q	56	58
COREQ6S1	P	96	96	B3Q6S1	Q	82	83	B5Q112S1	Q	46	47
B3Q7S1	P	64	65	B5Q113S1	Q	48	50				
B1Q1S1	D	76	77	B3Q8S1	P	45	45	B5Q114S1	D	65	66
B1Q2S1	D	93	93	B3Q9S1	P	57	58	B5Q12S1	D	59	62
B1Q3S1	D	90	90	B3Q10S1	D	16	16	B5Q13S1	Q	49	50
B1Q4S1	Q	85	85	B3Q11S1	P	64	66	B5Q14S1	Q	63	65
B1Q5S1	P	60	60	B3Q12S1	P	56	57				
B1Q6S1	P	68	68	B3Q13S1	P	77	78	B6Q1S1	P	96	96
B1Q7S1	Q	69	70	B3Q14S1	Q	49	51	B6Q2S1	Q	78	79
B1Q8S1	D	62	63	B3Q15S1	P	75	76	B6Q3S1	Q	83	84
B1Q9S1	Q	67	67	B6Q4S1	D	92	93				
B1Q10S1	P	92	93	B4Q1S1	P	98	99	B6Q5S1	Q	75	75
B1Q11S1	P	31	32	B4Q2S1	P	49	49	B6Q6S1	D	73	74
B1Q12S1	P	30	32	B4Q3S1	Q	87	87	B6Q7S1	P	79	81
B1Q13S1	D	89	89	B4Q4S1	D	52	54	B6Q8S1	P	71	73
B1Q14S1	Q	69	70	B4Q5_1S1	D	67	69	B6Q9S1	D	75	77
B1Q15S1	Q	84	85	B4Q5_2S1	Q	67	68	B6Q10S1	Q	44	45
B4Q6S1	P	63	63	B6Q11S1	D	73	74				
B2Q1S1	P	89	90	B4Q7S1	P	44	45	B6Q12S1	D	51	52
B2Q2S1	P	23	24	B4Q8S1	P	59	59	B6Q13S1	D	90	91
B2Q3S1	P	89	90	B4Q9S1	Q	74	75				
B2Q4S1	Q	71	71	B4Q10S1	Q	18	19	B7Q1S1	D	97	97
B2Q5S1	Q	33	34	B4Q11S1	Q	79	79	B7Q2S1	Q	86	87
B2Q6S1	P	59	59	B4Q121S1	D	93	94	B7Q3S1	D	94	95
B2Q7S1	P	18	18	B4Q122S1	D	90	91	B7Q4S1	D	79	81
B2Q8S1	D	51	53	B4Q123S1	D	93	93	B7Q5S1	Q	63	65
B2Q9S1	Q	85	86	B4Q124S1	D	96	96	B7Q6S1	Q	67	67
B2Q10S1	D	77	78	B4Q125S1	Q	90	90	B7Q7S1	D	48	50
B2Q111S1	D	93	94	B4Q126S1	Q	94	94	B7Q8S1	D	66	68
B2Q112S1	D	85	87	B7Q9S1	D	19	21				
B2Q113S1	D	97	98	B5Q1S1	P	94	94	B7Q10S1	P	79	81
B2Q114S1	D	96	97	B5Q2S1	P	95	95	B7Q11S1	P	42	43
B2Q115S1	D	72	72	B5Q3S1	P	82	83	B7Q12S1	P	61	63
B5Q4S1	P	32	34	B7Q13S1	P	98	98				
B5Q5S1	P	66	65	B7Q14S1	P	67	69				
B5Q6S1	P	47	48	B7Q15S1	P	85	85				

C. BLOCK LEVEL AVERAGE STATISTICS

This appendix contains the same information as appendix B but grouped according to the test block of the items.

Block #	Number Correct	Proportion Correct	R-Biserial	Reliability	Proportion Omit	Proportion notreached
1	10.51	.72	.55	.82	.04	.02
2	10.03	.70	.46	.81	.05	.04
3	9.31	.65	.53	.83	.05	.05
4	12.88	.74	.50	.83	.04	.03
5	10.63	.66	.57	.86	.05	.06
6	9.69	.77	.54	.81	.03	.03
7	10.26	.71	.55	.83	.05	.04

D. MEASUREMENT OF KEY SOCIAL FACTORS

One of the most challenging areas of the background questionnaire was the conceptualisation of the social variables such as the educational career of individuals, their social class, deprivation of the areas surveyed and whether they could be considered as urban or rural. This appendix describes how these variables were conceived.

Education

Respondents were asked to give their highest educational qualification and the number of years they had spent in continuous full-time education. This information was used to derive their position on the International Standard Classification of Education (ISCED) scale. For those who had qualifications these were also used to derive their level on the Scottish Qualifications Framework (SCQF) for weighting purposes.

The version of ISCED used was that which was in use at the time of IALS, to enable comparison with the earlier survey. This version of ISCED divided educational attainment into 7 categories, as follows:

ISCED 0 Education preceding the first level, usually begins at age 3, 4 or 5 and lasts one to three years (pre-primary). Includes those with no qualifications who finished fulltime continuous education before the age of 11.

ISCED 1 First level education, usually begins at age 5, 6 or 7 and lasts for about five or six years (primary). Includes those with no qualifications who finished fulltime continuous education between the ages of 11 and 14.

ISCED 2 Second level, first stage begins at about age 11 and lasts for about 3 years (lower secondary). Includes those with no qualifications who received full-time education until the age of 15 or 16.

ISCED 3 Second level second stage education begins at about age 14 or 15 and lasts for about three years (higher or upper secondary).

ISCED 4 Not used.

ISCED 5 Third level or higher education which leads to an award which is not equivalent to a university degree.

ISCED 6 Third level or higher education that leads to a university degree or equivalent.

ISCED 7 Third level or higher education post first degree that leads to a post-graduate university degree or equivalent.

ISCED 9 Education not definable by level.

The IALS report used the EUROSTAT version of the classification which meant that people whose highest qualification was Standard Grade or equivalent were coded as ISCED 2. The OECD version in use at the time coded these people as ISCED 3. Table A1 below shows how SCQF levels were matched to the version of ISCED used in the IALS survey to make it possible to describe the SSAL2009 participants on both scales.

It is critical to note that this does not imply an assumption that people with a certain number of years of schooling have specific qualifications—such an assumption is not tenable. This equivalence was used for attributing educational levels only for weighting. Substantive analysis was based on actual reported levels of qualification.

Table A1: ISCED and SCQF levels

ISCED in IALS	SCQF	
7	12	Professional Development Award, Doctorate
7	11	Professional Development Award, SVQ 5, Masters, Post Graduate Diploma, Post Graduate Certificate
6	10	Professional Development Award, Honours Degrees, Graduate Diploma
5	9	Professional Development Award, SVQ 4, Ordinary Degree, Graduate Certificate
5	8	Professional Development Award, SVQ4, Higher National Diploma, Diploma of Higher Education
5	7	Advanced higher, Professional Development Award, SVQ3, Higher National Certificate, Certificate of Higher Education
3	6	Higher, Professional Development Award, SVQ3, National Progression Award, National Certificates
2	5	Intermediate 2, Credit Standard Grade, (grades 1-2), National Progression Award, National Certificates
2	4	Intermediate 1, General Standard Grade (grades 3-4), National Progression Award, National Certificates
1	3	Access 3, Foundation Standard Grade (grades 5-6), National Progression Award, National Certificates
1	2	Access 2, National Progression Award
1	1	Access 1

Social class

Occupation details were collected for economically active and retired people. Unemployed people were asked about their last job and retired people about their main previous occupation.

Occupations were first coded according to the *International Standard Classification of Occupations (ISCO-88),* International Labour Organisation (ILO, 1991). This coding was necessary for the data to be scaled using the statistical models developed for IALS. These codes were then converted to those used in the *Standard Occupational Classification 2000*, ONS (HMSO, London 2001). For analysis and reporting of results in the SSAL2009 report, the SOC-2000 codes were used to derive social class according to the *National Statistics Socio-economic Classification (NS-SEC),* ONS (HMSO, London, 2005). Social class is presented in the SSAL2009 report in the following categories:

NS-SEC 1	Managerial and professional
NS-SEC 2	Intermediate occupations
NS-SEC 3	Small employers and own account workers
NS-SEC 4	Lower supervisory and technical occupations
NS-SEC 5	Semi-routine and routine occupations
Unclassified	Never worked or insufficient information available

Scottish Index of Multiple Deprivation

From http://www.scotland.gov.uk/Topics/Statistics/SIMD/Methodology

The Scottish Index of Multiple Deprivation (SIMD) 2006 defines small area concentrations of multiple deprivation across all of Scotland in a consistent way. The Scottish Index of Multiple Deprivation 2006 combines 37 indicators across 7 domains, namely: current income, employment, health, education, skills and training, housing, geographic access and crime. The overall index is a weighted sum of the seven domain scores. The weighting for each domain is based on the robustness of the data, the time lag between data collection and the production of the SIMD and the relative importance of the domain in measuring multiple deprivation. The domain weightings were subject to sensitivity analysis to assess the effects of any changes in weights on the overall index ranks.

The SIMD provides a *relative* measure of deprivation which means that the main output from SIMD - the SIMD ranks - can be used to compare data zones by providing a relative ranking from most deprived (rank 1) to least deprived (rank 6,505). The SIMD cannot be used to determine 'how much' more deprived one data zone is than another e.g. it is not possible to say that data zone X, ranked 50, is twice as deprived as data zone Y, ranked 100. However it is possible to say that X is more deprived than Y.

The SIMD can be used to identify Scotland's most deprived small areas on the overall index and each individual domain, commonly by applying a cut off such as

10%, 15% or 20%. The cut off should be informed by whether it aims to target areas with the very highest concentrations of deprivation or to be wider ranging.

Both the SIMD 2004 and SIMD 2006 are heavily based on the methodology developed by the Social Disadvantage Research Centre at the University of Oxford. In developing of an area-based measure of deprivation, the implementation of the Scottish Index of Multiple Deprivation is the Scottish Government's response to the August 2003 report 'Measuring Deprivation in Scotland : Developing a Long-Term Strategy'.

Rural and urban areas

The indicator applied to “area type” was the 6-fold Urban Rural classification. The Scottish Government 6-fold and 8-fold urban rural classifications are intended to provide a consistent way of defining urban and rural areas across Scotland. Two main criteria have been used to develop the SG 6-fold and 8-fold urban rural classifications: population as defined by the General Register Office for Scotland and accessibility based on drive time analysis to differentiate between accessible and remote areas in Scotland.

Text, and more information from,
http://scotland.gov.uk/Publications/2008/07/29152642/7

APPENDIX E: BACKGROUND QUESTIONNAIRE

This appendix shows the actual text of the questions in the background questionnaire. Instructions to the interviewers have been removed from this version of the background questionnaire. Questions are followed by the original IALS 1996 question code where the coding has been changed.

Now, I would like to ask a few questions about yourself.

Q7.[GENDER]: 1 Male 2 Female

Q8.[AGE]: Please tell me your age.

_____ Years old
08 – (DK/Refused)
09 – (Not stated)

Q9.[MARITALS]: What is your <u>current</u> marital status?

1 Single
2 Married (including Common-Law)
3 Widowed
4 Divorced, Separated, etc.
5 Other
8 (DK/Refused)
9 (Not stated)

Q10. [RTOHH]: What is your relationship to the head of household?

00 Head of Household
01 Spouse of Head of Household
02 Common law partner of HH
03 Son/daughter (including adopted)
04 Son-in-law/daughter-in-law
05 Grandchild
06 Parent
07 Parent-in-law
08 Brother/sister (including adopted)
09 Brother/sister in law
10 Lodger/Boarder
11 Roommate
12 Other
98 (DK/Refused)
99 (Not stated)

SECTION A

A1. Were you born in SCOTLAND? *[IALS: A1]*

1 Yes —> GO TO A3
2 No —> CONTINUE TO A2
8 (DK/Refused) —> CONTINUE TO A2
9 (Not stated) —> CONTINUE TO A2

A2. In which country were you born? *[IALS: A2]*

A3. During your lifetime, how many years of formal education have you completed beginning with the first year of primary school and not counting repeated years at the same level? (OPEN ENDED AND CODE THE NUMBER OF YEARS) *[IALS: A7]*

_____

01-90 Total years of schooling
00 No schooling /Never had formal education
98 (DK/Refused)
99 (Not stated)

A4. Do you have any qualifications from school or college, or connected with work or a government training scheme?

1 Yes —> CONTINUE TO A5.
2 No qualifications —> GO TO A6.
8 (DK/Refused) —> CONTINUE TO A5.
9 (Not stated) —> CONTINUE TO A5.

A5. What is your highest qualification of the followings shown on this card? Start at the top of the list.

1 Doctorate
2 MA, Post Graduate Diploma, Post Graduate Certificate, SVQ5
3 Honours Degree, Graduate Diploma
4 Ordinary Degree, Graduate Certificate, SVQ4
5 Higher National Diploma, Diploma of Higher Education, Diploma of Further Education, Foundation Degree
6 Advanced Higher, Higher National Certificate, Certificate of Higher Education, SVQ3
7 Higher, GCE A/S level
8 Intermediate 2, Credit Standard Grade, (grades 1-2), SVQ2, GCSE Grades A-C, GCE O level, CSE Grade 1
9 Intermediate 1, General Standard Grade (grades 3-4), SVQ1, GCSE Grades D-G, CSE Grades 2-5
10 Access 3, Foundation Standard Grade (grades 5-6)
11 Access 2
12 Access 1
13 Other (please specify:……)
98 (DK/Refused)
99 (Not stated)

A6. What was the main reason you stopped your schooling when you did? *(IALS: A12.)*

01 Still in school
02 Had enough education
03 Had to work / financial reasons
04 Wanted to work / wanted to learn a trade
05 Family reasons (help family business, illness at home, marriage, pregnancy etc.)
06 Did not like school
07 Did not do well in school / boredom
08 Personal illness / disability
09 School not available / not accessible
10 To join the military
11 Other
97 (Don't know)
98 (Refused)
99 (Not stated)

SECTION B
Linguistic Information

B1. What language did you FIRST speak as a child?
(IALS: B1L1-B1L2.)

B2. How would you rate your current writing skills in (ASK FOR THE LANGUAGE SPECIFIED IN B1)?
(IALS: B5L1.)

1 Cannot write in that language
2 Poor
3 Fair
4 Good
5 Very Good
0 (Not applicable)
8 (DK/Refused)
9 (Not stated)

B3. How old were you when you first started to learn English? *(IALS: B6.)*

_____
01-90 Age first learned language
97 Does not speak English
98 (DK/Refused)
99 (Not stated)

B4a. What languages including English do you speak well enough to conduct a conversation? And what other language? And what else? And …*(IALS: B13L 1-B13L6.)*

B5. What language do you speak most often at home? *(IALS: B14.)*

B6. What language do you speak most often at work or school? *(IALS: B15.)*

B7. What language do you speak most often during leisure activities? *(IALS: B16.)*

B8. In what language can you express yourself most easily? *(IALS: B17.)*

SECTION D
I would now like to talk to you about your employment status.

D1. What is your current work situation? Are you...? *(IALS: D1.)*

1	Employed	—> GO TO D4
2	Retired	—> CONTINUE
3	Unemployed / looking for work	—> CONTINUE
4	Student (including Work Programs)	—> CONTINUE
5	Homemaker	—> CONTINUE
6	Other	—> CONTINUE
8	(DK/ Refused)	—> CONTINUE
9	(Not stated)	—> CONTINUE

D2. Did you work at a job or business at any time in the past 12 months (regardless of the number of hours per week)?*(IALS: D2.)*

1	Yes	—> GO TO D4
2	No	—> CONTINUE
8	(DK / Refused)	—> CONTINUE
9	(Not stated)	—> CONTINUE

D3. When did you last work in a job in a company/business? *(IALS: D3.)*

_____

0000	Never worked	—> GO TO E1
9998	(DK / Refused)	—> CONTINUE
9999	(Not stated)	—> CONTINUE

D4. How many different employers have you had in the past 12 months? *(IALS: D4.)*

_____

01-90	Number of different employers
98	(DK / Refused)
99	(Not stated)

D5. Did/do you work mostly full time (that is more than 30 hours per week) or part time (that is less than 30 hours per week)? *(IALS: D5.)*

1	Full time	—> GO TO D7
2	Part time	—> CONTINUE
8	(DK / Refused)	—> GO TO D7
9	(Not stated)	—> GO TO D7

D6 Why did you work part-time? *(IALS: D6.)*

01	Own illness or disability
02	Child care responsibilities
03	Other personal or family responsibilities
04	Going to school/ taking training
05	Could only find a part time work
06	Did not want to work full time
07	Retired
08	Other
98	(DK /Refused)
99	(Not stated)

I would now like to ask you some questions about your (main) job, (that is the job you worked the most hours for).

D7 What did the firm/organisation you worked for mainly make or do (at the place where you worked)?

Write in: _____

98	(DK/Refused)
99	(Not stated)

D8 What was the industry? (ASK AND WRITE IN FULL DETAILS) *(IALS: ISICR.)*

Write in: _____

D9b. Looking at this card could you please tell me which category your main job fits the best?

D9. Can you please tell me exactly, what was your (main) job? *(IALS: ISCOR.)*

Write in: _____

D9a. What did you mainly do in your job?

_____

98 (DK/Refused)
99 (Not stated)

D10. In total, about how many persons are employed by this business at all locations in the UK? *(IALS: D10.)*

1 Less than 20
2 20 to 99
3 100 to 199
4 200 to 499
5 500 or over
8 (DK/Refused)
9 (Not stated)

D11. What was your status at this job? Was it as an ….? *(IALS: D11.)*

1 Employee without supervisory responsibilities
2 Employee with limited supervisory or management responsibilities (5 persons or less)
3 Employee with more extensive supervisory or management responsibilities (more than 5 persons)
4 Self-employed without employees
5 Self employed with employees
6 Family worker (unpaid)
8 (DK/Refused)
9 (Not stated)

D12. What type of job was this? Was or is this job a …?*(IALS: D12.)*

1 Permanent
2 Temporary
8 (DK/Refused)
9 (Not stated)

D13. How many hours per week do/did you usually work in your (main) job/business - please exclude meal breaks and overtime? *(IALS: D13.)*

_____
01 (Less than 1 hour)
02-96 Hours worked per week
97 (97 hours or more per week)
98 (DK/Refused)
99 (Not stated)

D14. During the past 12 months, how many weeks did you work (at all jobs including time off for holidays, maternity leave, illness and strike and industrial actions)? *(IALS: D14.)*

_____

01 (Less than a week)
02-52 Weeks worked in past 12 months
98 (DK/Refused)
99 (Not stated)

SECTION E
Reading and writing at work

E1. The following questions refer to the job at which you worked the most hours in the last 12 months. How often (do/did) you read or use information from each of the following as part of your main job? Would you say every day, a few times a week, once a week, less than once a week, rarely or never? *(IALS: E1A-E1G.)*

		Every day	A few times a week	Once a week	Less than once a week	Rarely or never	(DK/ Refused)	(Not stated)
a)	Letters or memos?	1	2	3	4	5	8	9
b)	Reports, articles, magazines or journals?	1	2	3	4	5	8	9
c)	Manuals or reference books, including catalogues?	1	2	3	4	5	8	9
d)	Diagrams or schematics?	1	2	3	4	5	8	9
e)	Bills, invoices, spreadsheets or budget tables?	1	2	3	4	5	8	9
f)	Material written in a language other than English?	1	2	3	4	5	8	9
g)	Directions or instructions for medicines, recipes, or other products?	1	2	3	4	5	8	9
h)	Read or use information from computers?	1	2	3	4	5	8	9

E2. How often (do/did) you write or fill out each of the following as part of your main job? *(IALS: E2A-E2D.)*

		Every day	A few times a week	Once a week	Less than once a week	Rarely or never	(DK/ Refused)	(Not stated)
a)	Letters or memos?	1	2	3	4	5	8	9
b)	Forms or things such as bills, invoices or budgets?	1	2	3	4	5	8	9
c)	Reports or articles?	1	2	3	4	5	8	9
d)	Estimates or technical specifications?	1	2	3	4	5	8	9

E3. In your main job, how often do you use arithmetic or mathematics (that is, adding, subtracting, multiplying or dividing) to:… ? *(IALS: E3A-E3B.)*

		Every day	A few times a week	Once a week	Less than once a week	Rarely or never	(DK/ Refused)	(Not stated)
a)	measure or estimate the size or weight of objects?	1	2	3	4	5	8	9
b)	calculate prices, costs or budgets?	1	2	3	4	5	8	9

E4. How would you rate your reading skills in English for your main job? Would you say they are ...? *(IALS: E4.)*

1 Excellent
2 Good
3 Moderate
4 Poor
7 (No opinion)
8 (DK/ Refused)
9 (Not stated)

E5. To what extent are your reading skills in English limiting your job opportunities - for example, advancement or getting another job? Are they…?*(IALS: E5.)*

1 Greatly limiting
2 Somewhat limiting
3 Not at all limiting?
8 (DK/ Refused)
9 (Not stated)

E6. How would you rate your writing skills in English for your main job? Would you say they are…? *(IALS: E6.)*

1 Excellent
2 Good
3 Moderate
4 Poor
7 (No opinion)
8 (DK/ Refused)
9 (Not stated)

E7. To what extent are your writing skills in English limiting your job opportunities - for example, advancement or getting another job? Are they…? *(IALS: E7.)*

1 Greatly limiting
2 Somewhat limiting
3 Not at all limiting?
8 (DK/ Refused)
9 (Not stated)

E8. How would you rate your mathematical skills for your main job? Would you say they are…? *(IALS: E8.)*

1 Excellent
2 Good
3 Moderate
4 Poor
7 (No opinion)
8 (DK/ Refused)
9 (Not stated)

E9. To what extent are your mathematical skills limiting your job opportunities - for example, advancement or getting another job? Are they…? *(IALS: E9.)*

1 Greatly limiting
2 Somewhat limiting
3 Not at all limiting?
8 (DK/ Refused)
9 (Not stated)

SECTION F
Adult education and training

The following questions will deal with any education or training which you may have taken in the past 12 months, since [INSERT ACTUAL MONTH OF INTERVIEWING] last year.

F1. During the last 12 months since [INSERT ACTUAL MONTH OF INTERVIEWING] last year did you receive any training or education including courses, private lessons, correspondence courses, workshops, on-the-job training, apprenticeship training, arts, crafts, recreation courses or any other training or education? *(IALS: F1.)*

1 Yes —> CONTINUE
2 No —> GO TO F15
8 (DK/Refused) —> GO TO F15
9 (Not stated) —> GO TO F15

F2. In total, how many courses did you take in the past 12 months? *(IALS: F2.)*

_____
98 (DK/Refused)
99 (Not stated)

F3A_1. Here is a list of types of course. Which best describes the first of these courses? *(IALS: F3M1)*

01 Skill training/upgrading related to your occupation or profession
02 Skill training/upgrading not related to your occupation or profession
03 Personal development and communication skills (e.g. time management, team leadership, stress management)
04 Management/organisation training and development (include Human resource management)
05 Computer software training (e.g. Excel, Pagemaker)
06 Health and safety (e.g. first aid, lifting and handling goods, risk assessment)
07 Languages (business and conversational)
08 Operation and/or maintenance of machinery/equipment
09 Quality assurance/control
10 Sport and physical fitness
11 Other recreational activities (e.g. Bridge or painting)
12 Preliminary education and training
13 Humanities, Arts, Music and Social Studies
14 Science, Environment, Engineering, Technology or medicine
15 Business/legal studies
16 None of these
98 (DK/Refused)
99 (Not stated)

F3B_1. Which of these subject areas was this training or education? (SHOW CARD I)

01 Education, recreational and counselling services
02 Fine and applied arts
03 Humanities and related fields
04 Social sciences and related fields
05 Commerce, management and business administration
06 Agricultural and biological sciences and technologies
07 Engineering and applied sciences
08 Engineering and applied sciences, technologies and trades
09 Health professions, sciences and technologies
10 Mathematics and physical sciences
11 No specialisation
12 Other
98 (DK/Refused)
99 (Not stated)

F5_1. Were you taking this training or education towards... *(IALS: F5M1)*

1 A university degree/ diploma/certificate
2 A college diploma/ certificate
3 A trade-vocational diploma/certificate
4 An apprenticeship certificate
5 An elementary or secondary school diploma
6 Professional or career upgrading
7 Other
8 (DK/Refused)
9 (Not stated)

F6_1. Was this training or education given by...(SHOW CARD K - CODE ALL THAT APPLY) *(IALS: F6AM1-F6GM1.)*

		Yes	No	(DK/ Refused)	(Not stated)
a)	A University (including Open University) or other higher education establishment	1	2	8	9
b)	A further education college or other adult education college/centre	1	2	8	9
c)	A commercial organisation (for example, a private training provider)	1	2	8	9
d)	A producer or supplier of equipment	1	2	8	9
e)	A non-profit organisation such as an employer association, voluntary organisation or a trade union	1	2	8	9
f)	An employer or a parent company	1	2	8	9
g)	Other government or local authority organization	1	2	8	9
h)	Other provider	1	2	8	9

F8_1. For how many weeks did this training or education last? *(IALS: F8M1)*

F9_1. On average, how many days per week was it? *(IALS: F9M1)*

F10_1. On average, how many hours per day was it? *(IALS: F10M1)*

			(Less than 1)	(DK/ Refused)	(Not stated)
F8_1	Number of weeks	_____	1	98	99
F9_1	Number of days	_____	1	98	99
F10_1	Number of hours per day	_____	1	98	99

F11_1. What was the main reason you took this training or education? Was it for...? *(IALS: F11M1)*

1 Career/job related purposes —> CONTINUE
2 Personal interest? —> GO TO F3A_2
3 (Other) —> GO TO F3A_2
8 (DK/Refused) —> GO TO F3A_2
9 (Not stated) —> GO TO F3A_2

F12_1. To what extent are you using the skills or knowledge acquired in this training at work? *(IALS: F12M1)*

1 To a great extent
2 Somewhat
3 Very little
4 Not at all
7 (Not applicable)
8 (DK/Refused)
9 (Not stated)

[Questions repeated for second and third courses]

F15. In the last 12 months, since [INSERT ACTUAL MONTH last year] was there any training course or education that you WANTED to take for career or job-related reasons but did not? *(IALS: F15.)*

1	Yes	—> CONTINUE
2	No	—> GO TO Q. F17
8	(DK/Refused)	—> GO TO Q. F17
9	(Not stated)	—> GO TO Q. F17

F16. What were the reasons you did not take this training or education? *(IALS: F16A-F16K.)*

	Yes	No	(DK/ Refused)	(Not stated)
a) Too busy/lack of time	1	2	8	9
b) Too busy at work	1	2	8	9
c) Course not offered	1	2	8	9
d) Family responsibilities	1	2	8	9
e) Too expensive/no money	1	2	8	9
f) Lack of qualifications	1	2	8	9
g) Lack of employer support	1	2	8	9
h) Course offered at inconvenient time	1	2	8	9
i) Language reasons	1	2	8	9
j) Health reasons	1	2	8	9
k) Other	1	2	8	9

F17. In the last 12 months was there any other training course that you WANTED to take but did not, such as hobby, recreational or interest courses?*(IALS: F17.)*

1	Yes	—> CONTINUE
2	No	—> GO TO SECTION 'Reading and writing at home' (G1)
8	(DK/Refused)	—> GO TO SECTION 'Reading and writing at home' (G1)
9	(Not stated)	—> GO TO SECTION 'Reading and writing at home' (G1)

F18. What were the reasons you did not take this training or education? *(IALS: F18A-F18K.)*

	Yes	No	(DK/ Refused)	(Not stated)
a) Too busy/lack of time	1	2	8	9
b) Too busy at work	1	2	8	9
c) Course not offered	1	2	8	9
d) Family responsibilities	1	2	8	9
e) Too expensive/no money	1	2	8	9
f) Lack of qualifications	1	2	8	9
g) Lack of employer support	1	2	8	9
h) Course offered at inconvenient time	1	2	8	9
i) Language reasons	1	2	8	9
j) Health reasons	1	2	8	9
k) Other	1	2	8	9

Section G
Reading and writing at home

The next few questions deal with reading and writing in your daily life excluding work (or school).

G1. I am going to read you a list of activities. Please tell me if you do each of them daily, weekly, every month, several times a year or never? How often do you... *(IALS: G1A-G1H)*

		Daily	Weekly	Monthly	Several times a year	Never	(DK/Refused)	(Not stated)
a)	Use a public library?	1	2	3	4	5	8	9
b)	Attend a movie, play or concert?	1	2	3	4	5	8	9
c)	Attend or take part in a sporting event?	1	2	3	4	5	8	9
d)	Write letters or anything else that is more than one page in length?	1	2	3	4	5	8	9
e)	Participate in voluntary or community organizations?	1	2	3	4	5	8	9
f)	Reading newspapers or magazines?	1	2	3	4	5	8	9
g)	Reading books?	1	2	3	4	5	8	9
h)	Listening to radio, records, tapes, cassettes, or compact discs?	1	2	3	4	5	8	9
i)	Use a personal computer/laptop/ PDA?	1	2	3	4	5	8	9

G2. Do you ever do any of these activities in a language other than English? *(IALS: G2)*

1	Yes	—> CONTINUE
2	No	—> GO TO G4
8	(DK/Refused)	—> GO TO G4
9	(Not stated)	—> GO TO G4

G3. Which of the following activities have you ever done in a language other than English? *(IALS: G3A-G3H)*

		Yes	No	(DK/ Refused)	(Not stated)
a)	Use a public library	1	2	8	9
b)	Attend a movie, play or concert	1	2	8	9
c)	Attend or take part in a sporting event	1	2	8	9
d)	Write letters or anything else that is more than one page in length	1	2	8	9
e)	Participate in voluntary or community organizations	1	2	8	9
f)	Reading newspapers or magazines	1	2	8	9
g)	Reading books	1	2	8	9
h)	Listening to radio, records, tapes, cassettes, or compact discs, Ipods or mp3 players	1	2	8	9

G4. How much time do you usually spend each day watching television or videos (DVDs)? *(IALS: G4)*

1 Not on a daily basis
2 1 hour or less per day
3 1 to 2 hours per day
4 More than 2 hours but less than 5
5 5 or more hours per day
6 Do not have a television or video
8 (DK/Refused)
9 (Not stated)

G5. Do you ever watch television or videos in a language other than English? *(IALS: G5)*

1 Yes
2 No
8 (DK/Refused)
9 (Not stated)

G6. Which of the following materials do you currently have in your home? *(IALS: G6A-G6E)*

	Yes	No	(DK/ Refused)	(Not stated)
a) Daily newspaper	1	2	8	9
b) Weekly newspaper/magazines	1	2	8	9
c) More than 25 books	1	2	8	9
d) A (multi-volume) encyclopedia	1	2	8	9
e) A dictionary	1	2	8	9

G7. I am now going to show you a list of different parts of a newspaper. Please tell me which parts you generally read when looking at a newspaper. *(IALS: G7A-G7O)*

	Yes	No	(Respondent does not read a newspaper)	(DK/ Refused)	(Not stated)
a) Classified ads	1	2	3	8	9
b) Other advertisements	1	2	3	8	9
c) National/international news	1	2	3	8	9
d) Regional or local news	1	2	3	8	9
e) Sports	1	2	3	8	9
f) Home, fashion or health	1	2	3	8	9
g) Editorial page	1	2	3	8	9
h) Financial news or share listings	1	2	3	8	9
i) Comics	1	2	3	8	9
j) TV listings	1	2	3	8	9
k) Movie or concert listings	1	2	3	8	9
l) Book, movie or art reviews	1	2	3	8	9
m) Horoscope/stars	1	2	3	8	9
n) Advice column	1	2	3	8	9
o) Special interest sections, e.g. education, motor or computer section	1	2	3	8	9
p) Personal finance, money advice	1	2	3	8	9
r) Other	1	2	3	8	9

G8. Would you say you follow what's going on in current events, government and public affairs...? *(IALS: G8)*

1 Most of the time
2 Some of the time
3 Only now and then
4 Hardly at all
8 (DK/Refused)
9 (Not stated)

G9. I would like to know how you usually get information about current events, public affairs, and the government. How much information do you get from... *(IALS: G9A-G9E)*

		A lot	Some	Very little	None	(DK/ Refused)	(Not stated)
a)	Newspapers	1	2	3	4	8	9
b)	Magazines	1	2	3	4	8	9
c)	Radio	1	2	3	4	8	9
d)	Television	1	2	3	4	8	9
e)	Family members, friends or co-workers	1	2	3	4	8	9
f)	Internet	1	2	3	4	8	9

G10. Sometimes people need help from family members or friends to read and write (in English). How often do you need help from others with... *(IALS: G10A-G10G)*

		Often	Sometimes	Never	(DK/ Refused)	(Not stated)
a)	Reading newspaper articles?	1	2	3	8	9
b)	Reading information from government departments, businesses or other institutions?	1	2	3	8	9
c)	Filling out forms such as applications or bank deposit slips?	1	2	3	8	9
d)	Reading instructions such as on medicine bottles?	1	2	3	8	9
e)	Reading instructions on 'packaged' goods in shops/stores or supermarkets?	1	2	3	8	9
f)	Doing basic arithmetic, that is, adding, subtracting, multiplying and dividing?	1	2	3	8	9
g)	Writing notes and letters?	1	2	3	8	9

G11. How would you rate your reading skills in English needed in daily life? (SHOW CARD S)

G12. And how would you rate your writing skills in English needed in daily life? (SHOW CARD S)

G13. And how would you rate your mathematical skills in English needed in daily life? *(IALS: G11-G13)*

	G11.	G12.	G13.
Excellent	1	1	1
Good	2	2	2
Moderate	3	3	3
Poor	4	4	4
(No opinion)	7	7	7
(DK/Refused)	8	8	8
(Not stated)	9	9	9

G14. All things considered how satisfied are you with your reading and writing skills in English? Are you …?*(IALS: G14)*

1 Very satisfied
2 Somewhat satisfied
3 Somewhat dissatisfied
4 Very dissatisfied
7 (No opinion)
8 (DK/Refused)
9 (Not stated)

G15. Have you ever had …

G16. Did you have this issue while you were in primary or secondary school?

G17. Do you have this issue now?
(IALS: G15A-E, G16A-E, G17A-E)

	Q15. EVER HAD	Yes	No	(DK/Refused)	(Not stated)	**Q16. IN SCHOOL**	Yes	No	(DK/Refused)	(Not stated)	**Q17. NOW**	Yes	No	(DK/Refused)	(Not stated)
A.	Eye/visual trouble of the kind that is not corrected by glasses	1	2	8	9	→	1	2	8	9		1	2	8	9
B.	Hearing problems	1	2	8	9	→	1	2	8	9		1	2	8	9
C.	Speech disability	1	2	8	9	→	1	2	8	9		1	2	8	9
D.	Learning disability	1	2	8	9	→	1	2	8	9		1	2	8	9
E.	Other disability or health problem for 6 months or more	1	2	8	9	→	1	2	8	9		1	2	8	9

Section H
Family literacy

H1. Are you the parent or guardian of any children aged 6 to 18 that are presently living with you?
(IALS: H1)

1 Yes —> CONTINUE
2 No —> GO TO SECTION 'Parental Information' (C1)
8 (DK/Refused) —> GO TO SECTION 'Parental Information' (C1)
9 (Not stated) —> GO TO SECTION 'Parental Information' (C1)

H1A. How many children aged 6 to 18 are presently living with you?

_____
98 (DK/Refused)
99 (Not stated)

H2. What is the age of the youngest child presently living with you, aged 6 to 18? *(IALS: H2)*

_____

98 (DK/Refused)
99 (Not stated)

H2A. What is the name of the youngest child presently living with you, aged 6 to 18? (RECORD NAME)

_____

98 (DK/Refused)
99 (Not stated)

H3A. What is the highest grade of schooling that this child (INSERT NAME OF YOUNGEST CHILD) completed? (not counting repeated years at the same level)? *(IALS: H3A)*

0 Primary but has not completed yet
1 Primary
2 Secondary
3 Post-secondary
4 Trade vocational
5 No schooling
8 (DK/Refused)
9 (Not stated)

H3B. What is the highest year of schooling that (INSERT NAME OF YOUNGEST CHILD) completed (not counting repeated years at the same level)? *(IALS: H3B)*

0 Primary but has not completed a full year yet
1 1 year (P1)
2 2 years (P2)
3 3 years (P3)
4 4 years (P4)
5 5 years (P5)
6 6 years (P6)
7 7 years (P7)
8 8 years (S1)
9 9 years (S2)
10 10 years (S3)
11 11 years (S4)
12 12 years (S5)
13 13 years (S6)
14 No schooling

98 (DK/Refused)
99 (Not stated)

H4. How often would you say (INSERT NAME OF YOUNGEST CHILD) reads for pleasure? Would you say... *(IALS: H4)*

1 Every day
2 A few times a week
3 Several times a month
4 A few times a month
5 Once a month or less
6 (Never)
7 (Not applicable (child can't read))
8 (DK/Refused)
9 (Not stated)

H6. Given (INSERT NAME OF YOUNGEST CHILD)'s age, how satisfied are you with the way he/she reads? Would you say that you are…? *(IALS: H6)*

1 Very satisfied
2 Somewhat satisfied
3 Somewhat dissatisfied
4 Very dissatisfied
7 (No opinion)
8 (DK/Refused)
9 (Not stated)

H7. Could you please tell me if each of the following statements are true or false of your household? *(IALS: H7A–H7H)*

	Statements	True	False	(DK/ Refused)	(Not stated)
a)	There is a variety of books in your home	1	2	8	9
b)	There is a variety of magazines and other reading material in your home	1	2	8	9
c)	Your children often see you or your partner/spouse reading.	1	2	8	9
d)	Your children learned to read before starting school.	1	2	8	9
e)	Your children have a certain amount of time set aside each day for reading at home.	1	2	8	9
f)	Your children are limited in the amount of time you allow them to watch TV.	1	2	8	9
g)	Your children often choose the books they read	1	2	8	9
h)	Your children have their own books and place to keep them	1	2	8	9

H8. I would like to read you a list of different things that some parents may find useful in helping their children to become good readers. For each one, I would like you to tell me whether you think it is very useful, somewhat useful, not very useful or not at all useful. *(IALS: H8A–H8E)*

		Very useful	Somewhat useful	Not very useful	Not at all useful	(DK/Refused)	(Not stated)
a)	Reading lists supplied to parents by schools and other educational experts.	1	2	3	4	8	9
b)	Materials such as games, activities and books being supplied to parents by schools or other educational experts that would help parents encourage their children to read	1	2	3	4	8	9
c)	A close parent-teacher relationship	1	2	3	4	8	9
d)	Schools providing parents with help in understanding assessments of a child's reading abilities.	1	2	3	4	8	9
e)	Access for children to books, through either public or school libraries.	1	2	3	4	8	9

SECTION C
Parental Information

C2. What was the highest level of schooling that your mother (female guardian) completed? *(IALS: C5)*

10 No Schooling
2 Below Primary School
3 Completed Primary School (current leaving age 11)
4 Secondary School (current leaving age 16, GCSE or equivalent)
5 Upper Secondary School (current leaving age 18, GCE A-level or equivalent)
6 Higher/Further Education NOT leading to a degree (eg. HNC/HND)
7 University/College first degree
8 Postgraduate University degree
9 Education not definable by level
98 (DK/Refused)
99 (Not stated)

C4. What was the highest level of schooling that your father (male guardian) completed? *(IALS: C11)*

10 No Schooling
2 Below Primary School
3 Completed Primary School (current leaving age 11)
4 Secondary School (current leaving age 16, GCSE or equivalent)
5 Upper Secondary School (current leaving age 18, GCE A-level or equivalent)
6 Higher/Further Education NOT leading to a degree (eg. HNC/HND)
7 University/College first degree
8 Postgraduate University degree
9 Education not definable by level
98 (DK/Refused)
99 (Not stated)

SECTION I
Immigration

I2. Which year did you first come to live in this country? *(IALS: A3)*

_____
9998 (DK/Refused)
9999 (Not stated)

I3. In total how many years have you lived in Scotland? *(IALS: A4)*

_____
1 (Less than one year)
98 (DK/Refused)
99 (Not stated)

I4. Before you first came to live in Scotland in (YEAR IN I2), what was the highest level of schooling you had completed? *(IALS: A5)*

10 No Schooling
2 Below Primary School
3 Completed Primary School (current leaving age 11)
4 Secondary School (current leaving age 16, GCSE or equivalent)
5 Upper Secondary School (current leaving age 18, GCE A-level or equivalent)
6 Higher/Further Education NOT leading to a degree (eg. HNC/HND)
7 University/College first degree
8 Postgraduate University degree
9 Education not definable by level
98 (DK/Refused)
99 (Not stated)

I5. May I just check, to which of these groups do you consider you belong?

[Standard UK census groups offered]

Income

Finally, I would like to get a bit of general household information.

INC1. From which of the following sources did you receive income in 2008? *(IALS: J1A-J1E)*

		Yes	No	(DK/Refused)	(Not stated)
a)	Wages, salary or self-employment	1	2	8	9
b)	Government or State sources (social assistance, unemployment support, excluding pensions)	1	2	8	9
c)	Government or State pensions (old age, disability)	1	2	8	9
d)	Interest, dividends, investment, property or private pensions	1	2	8	9
e)	Other sources such as alimony, scholarship, child support, etc.	1	2	8	9

INC2. What is the best estimate of your personal income in 2008 from all sources, including those just mentioned? *(IALS: J2Q)*

0 No income —> Go to INC4.
1 Up to £9,500
2 £9,501-£15,000
3 £15,001-£20,000
4 £20,001-£29,500
5 £29,501 or more
8 (DK/Refused)
9 (Not stated)

INC3. What is the best estimate of your personal income from only wages, salary or self-employment in 2008? *(IALS: J3Q)*

0 No income —> Go to INC4.
1 Up to £9,500
2 £9,501-£15,000
3 £15,001-£20,000
4 £20,001-£29,500
5 £29,501 or more
8 (DK/Refused)
9 (Not stated)

INC4. Including yourself how many people live in this household? *(IALS: J4)*

1 1 person only —> GO TO INTERVIEWER SECTION
2 2 people
3 3 people
4 4 people
5 5 people
6 6 people
7 7 or more people
8 (DK/Refused)
9 (Not stated)

INC5. What is the best estimate of the total income of all household members (including yourself) from all sources in 2008? *(IALS: J5Q)*

0	No income
1	Up to £11,500
2	£11,501-£17,500
3	£17,501-£28,000
4	£28,001-£44,000
5	£44,001 or more
8	(DK/ Refused)
9	(Not stated)

APPENDIX F: SAMPLE QUESTIONS FROM THE LITERACY TASK BOOKLETS

The following pages contain four examples of stimuli and questions used in SSAL2009 2009 (and in IALS in 1996). They come from blocks 2 and 6 of the matrix sampling design. As printed in the test booklets the stimulus would have been on the left hand page with the questions on the facing (right hand) page. Short explanations of each question are provided, derived from the Technical Report for the 1996 IALS survey.

Ingredients for 4 people:

1 garlic clove
1 onion
3 tablespoons oil
500 grams of fresh red tomatoes or 500 gram can of tomatoes
salt
1 teaspoon sugar
6 eggs

Scrambled Eggs with Tomatoes

Fry chopped garlic and onion in frying pan with oil until transparent. Add tomatoes that have been peeled and chopped (if they are fresh) or mashed with a fork (if they are canned). Add salt and sugar to lessen the acidity. When the mixture begins to thicken, add the eggs, already beaten, and stir well with a wooden spoon. Cook until eggs are set.

Questions 1-3. Use the recipe for scrambled eggs on the opposite page to answer questions 1 to 3.

1. Why does the recipe call for sugar?

   _____

2. If you want to make enough scrambled eggs for six people, how many eggs should you use?

   _____

3. If you decide to make just enough scrambled eggs for two people, how many tablespoons of oil would you need?

   _____

Description

Respondents are asked to use the recipe for scrambled eggs with tomatoes, which gives the ingredients required for four people. The first question is at Level 1 on the Prose scale, with the reader asked to locate the sentence explaining why the recipe includes sugar.

The second question is more complex, requiring calculation of how many eggs are needed for six people rather than four. Here they must know how to calculate or determine the ratio needed as well as locate the original number of eggs. This is Level 3 on the Quantitative scale.

The third question is at Level 2 on the Quantitative Scale because a large proportion of respondents found it easier to halve an ingredient rather than multiply it by 1.5 as required in question 2.

IMPATIENS

Like many other cultured plants, impatiens plants have a long history behind them. One of the older varieties was sure to be found on grandmother's windowsill. Nowadays, the hybrids are used in many ways in the house and garden.

Origin: The ancestors of the impatiens, *Impatiens sultani* and *Impatiens holstii*, are probably still to be found in the mountain forests of tropical East Africa and on the islands off the coast, mainly Zanzibar. The cultivated European plant received the name *Impatiens walleriana.*

Appearance: It is a herbaceous bushy plant with a height of 30 to 40 cm. The thick, fleshy stems are branched and very juicy, which means, because of the tropical origin, that the plant is sensitive to cold. The light green or white speckled leaves are pointed, elliptical, and slightly indented on the edges. The smooth leaf surfaces and the stems indicate a great need of water.

Bloom: The flowers, which come in all shades of red, appear plentifully all year long, except for the darkest months. They grow from "suckers" (in the stem's "armpit").

Assortment: Some are compact and low-growing types, about 20 to 25 cm. high, suitable for growing in pots. A variety of hybrids can be grown in pots, window boxes, or flower beds. Older varieties with taller stems add dramatic colour to flower beds.

General care: In summer, a place in the shade without direct sunlight is best; in fall and spring, half-shade is best. When placed in a bright spot during winter, the plant requires temperatures of at least 20°C; in a darker spot, a temperature of 15°C will do. When the plant is exposed to temperatures of 12-14°C, it loses its leaves and won't bloom anymore. In wet ground, the stems will rot.

Watering: The warmer and lighter the plant's location, the more water it needs. Always use water without a lot of minerals. It is not known for sure whether or not the plant needs humid air. In any case, do not spray water directly onto the leaves, which causes stains.

Feeding: Feed weekly during the growing period from March to September.

Repotting: If necessary, repot in the spring or in the summer in light soil with humus (prepacked potting soil). It is better to throw the old plants away and start cultivating new ones.

Propagating: Slip or use seeds. Seeds will germinate in ten days.

Diseases: In summer, too much sun makes the plant woody. If the air is too dry, small white flies or aphids may appear.

Questions 1-3. Use the article about the flower impatiens on the opposite page to answer questions 1 to 3.

1. According to the article, what do the smooth leaf surface and the stems suggest about the plant?

_____

_____

2. Using the information in the article, list two reasons why impatiens might be considered good plants to have.

_____

_____

3. What happens when the impatiens plant is exposed to temperatures of 14˚C or below?

_____

_____

Description

Questions 1 and 3 are at Prose Level 2. Like Prose Level 1 questions, respondents have to locate information, but there are more varied demands in terms of the number of items of information required or the distracting information. In both of these questions there is distracting information in the sentence immediately preceding the sentence with the correct response.

Compound Interest
Compounded Annually

Principal	Period	4%	5%	6%	7%	8%	9%	10%	12%	14%	16%
£100	1 day	0.011	0.014	0.016	0.019	0.022	0.025	0.027	0.033	0.038	0.044
	1 week	0.077	0.096	0.115	0.134	0.153	0.173	0.192	0.230	0.268	0.307
	6 mos	2.00	2.50	3.00	3.50	4.00	4.50	5.00	6.00	7.00	8.00
	1 year	4.00	5.00	6.00	7.00	8.00	9.00	10.00	12.00	14.00	16.00
	2 years	8.16	10.25	12.36	14.49	16.64	18.81	21.00	25.44	29.96	34.56
	3 years	12.49	15.76	19.10	22.50	25.97	29.50	33.10	40.49	48.15	56.09
	4 years	16.99	21.55	26.25	31.08	36.05	41.16	46.41	57.35	68.90	81.06
	5 years	21.67	27.63	33.82	40.26	46.93	53.86	61.05	76.23	92.54	110.03
	6 years	26.53	34.01	41.85	50.07	58.69	67.71	77.16	97.38	119.50	143.64
	7 years	31.59	40.71	50.36	60.58	71.38	82.80	94.87	121.07	150.23	182.62
	8 years	36.86	47.75	59.38	71.82	85.09	99.26	114.36	147.60	185.26	227.84
	9 years	42.33	55.13	68.95	83.85	99.90	117.19	135.79	177.31	225.19	280.30
	10 years	48.02	62.89	79.08	96.72	115.89	136.74	159.37	210.58	270.72	341.14
	12 years	60.10	79.59	101.22	125.22	151.82	181.27	213.84	289.60	381.79	493.60
	15 years	80.09	107.89	139.66	175.90	217.22	264.25	317.72	447.36	613.79	826.55
	20 years	119.11	165.33	220.71	286.97	366.10	460.44	572.75	864.63	1,274.35	1,846.08

Questions 9-11. Use the table giving amounts of compound interest on the opposite page to answer questions 9 to 11.

9. You wish to invest £100 for 20 years. List all the rates on the table that will yield more than £500 in interest.

_____

10. Using the information in the table, calculate the total amount of money you will have if you invest £100 at a rate of 6% for 10 years.

_____

11. If you wanted to more than double your principal within five years, what rate of interest on this table would you need?

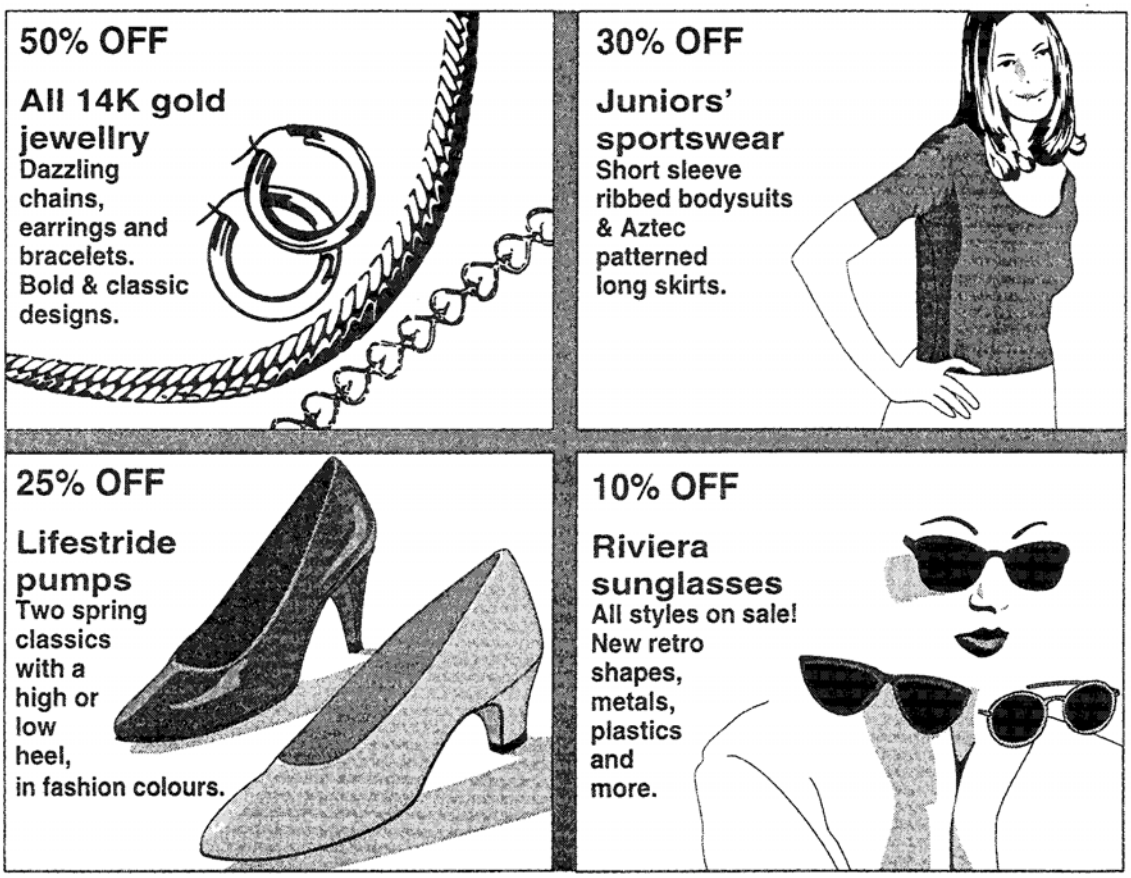
_____

Description

Question 9 is at Document Level 2, requiring respondents simply to identify the correct information in the table.

Question 11 is at Document Level 3 because it requires the respondent to identify a single interest rate that meets specified criteria. The respondent also needs to understand that “doubling principal” means looking for interest greater than £100.

Question 10 is at Quantitative Level 4. It requires respondents to find the appropriate interest rate and to perform a simple calculation of a 3 digit and a 5 digit number including a decimal point.


50% OFF
All 14K gold jewellry
Dazzling chains, earrings and bracelets. Bold & classic designs.
30% OFF
Juniors' sportswear
Short sleeve ribbed bodysuits & Aztec patterned long skirts.
25% OFF
Lifestride pumps
Two spring classics with a high or low heel, in fashion colours.
10% OFF
Riviera sunglasses
All styles on sale! New retro shapes, metals, plastics and more.

Questions 4-5. Use the advertisement for women's clothes and accessories on the opposite page to answer questions 4 and 5.

4. Calculate the total amount you would pay on sale for a gold chain originally priced at £125 plus matching earrings originally priced at £79.

_____

5. You buy a bodysuit that originally cost £45 and a pair of shoes that originally cost £69. Calculate the total amount you will save on the two items because of the sale.

_____

Description

Question 4 is at Quantitative Level 3 requires respondents to locate the percentage discount for the two items, calculate the discounted prices for the items and add them.

Question 5 is at Quantitative Level 5 and is one of the most challenging items on the Quantitative scale. Respondents have to work out what calculation is required and then follow it through without error.