1352.0.55.094

Australian Bureau of Statistics

Research Paper

# Experimental Estimates of Adult Literacy for Local Government Areas

www.abs.gov.au

New Issue

# Research Paper

# Experimental Estimates of Adult Literacy for Local Government Areas

Pramod Adhikari

National Centre for Education and Training Statistics

Methodology Advisory Committee

13 June 2008, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) THURS 21 AUG 2008

ABS Catalogue no. 1352.0.55.094



Produced by the Australian Bureau of Statistics

## INQUIRIES

The ABS welcomes comments on the research presented in this paper. For further information, please contact Dr Pramod Adhikari, National Centre for Education and Training Statistics on Canberra (02) 6252 7646 or email <analytical.services@abs.gov.au>

# CONTENTS


ABSTRACT . . . . . 1

1. INTRODUCTION . . . . . 1

2. THE ADULT LITERACY AND LIFE SKILLS SURVEY 2006 . . . . . 3

3. MEASUREMENT OF LITERACY . . . . . 5

4. ESTIMATION METHOD . . . . . 6
   - 4.1 Variables . . . . . 6
   - 4.2 Model . . . . . 10
   - 4.3 Results from the random intercept logistic model . . . . . 11

5. PREDICTION METHOD . . . . . 17
   - 5.1 Variance estimation . . . . . 18
   - 5.2 Predicted results . . . . . 19
   - 5.3 Comparison with other measures of disadvantage . . . . . 20

6. CONCLUSION . . . . . 24

ACKNOWLEDGEMENTS . . . . . 24

REFERENCES . . . . . 25

APPENDIX . . . . . 27

# EXPERIMENTAL ESTIMATES OF ADULT LITERACY FOR LOCAL GOVERNMENT AREAS

Pramod Adhikari
National Centre for Education and Training Statistics

## ABSTRACT

Large national surveys such as the *Survey of Aspects of Literacy* and the *Adult Literacy and Life Skills Survey* are able to provide literacy estimates for national and state levels. However, due to sample size constraints, it is not possible to produce estimates for smaller geographical areas using the sample data alone. The purpose of this paper is to derive experimental estimates of adult literacy for Local Government Areas from the Adult Literacy and Life Skills Survey 2006 (ALLS 2006). The paper uses a small area estimation technique – specifically a multilevel random intercept model – to derive estimates for small geographical areas.

## 1. INTRODUCTION

Education, especially one's literacy level is very important in a person's social and economic life. Literacy is now seen as how adults use written information to function in society. In 1992, the Organisation for Economic Cooperation and Development (OECD) concluded that low literacy levels were a serious threat to economic performance and social cohesion (OECD, 1992). Literacy is more complex than the ability to read. The *International Adult Literacy Survey* (IALS) defines literacy as an "adult behaviour by which adults, using printed and written information, are better able to function in society, to achieve one's goals, and to develop one's knowledge and potential" (OECD and Statistics Canada, 1995, p. 14).

In 1996, Australia took part in the *International Adult Literacy Survey* (IALS) which was coordinated by the OECD and Statistics Canada. This international study involved many countries undertaking similar surveys over a four year period, enabling international comparisons of various aspects of literacy measured by the IALS. The Australian component of IALS 1996 is known as the *Survey of Aspects of Literacy* (SAL).

A decade later, in 2006, Australia conducted the *Adult Literacy and Life Skills Survey* (ALLS 2006) as part of a further international study coordinated by Statistics Canada and the OECD.

The ALLS 2006 provides information on knowledge and skills in the following four domains:

1. *Prose literacy*: the ability to understand and use information from various kinds of narrative texts, including texts from newspapers, magazines and brochures;
2. *Document literacy*: the knowledge and skills required to locate and use information contained in various formats, including job applications, payroll forms, transportation schedules, maps, tables and charts;
3. *Numeracy*: the knowledge and skills required to effectively manage and respond to the mathematical demands of diverse situations; and
4. *Problem solving*: goal-directed thinking and action in situations for which no routine solution is available.

Neither the SAL nor the ALLS define literacy in terms of a basic threshold – above which someone is 'literate' and below which someone is 'illiterate'. For each literacy domain, proficiency is measured on a scale ranging from 0 to 500 points. To facilitate analysis, these continuous scores have been grouped into five skill levels (only four levels were defined for the problem solving scale) with Level 1 being the lowest measured level of literacy. To assist with interpreting the results, Level 3 is regarded by the survey developers as the "minimum required for individuals to meet the complex demands of everyday life and work in the emerging knowledge-based economy" (Statistics Canada and OECD, 2005).

In the sections that follow, we provide a description of the ALLS 2006 and how literacy levels for individuals have been measured. We then describe the methods used to estimate the prevalence of 'low literacy' at the Local Government Area level. We present our results and provide some concluding remarks.

## 2. THE ADULT LITERACY AND LIFE SKILLS SURVEY 2006

The conduct of the ALLS 2006 in Australia was jointly funded by the Department of Education Science and Training (DEST), the Department of Employment and Workplace Relations (DEWR) and the Australian Bureau of Statistics (ABS). Other countries that have participated, or are currently participating in the study include the United States of America, Bermuda, Canada, Italy, Mexico, Norway, Switzerland, Hungary, the Netherlands, New Zealand and South Korea.

The ALLS 2006 is designed to identify and measure literacy, numeracy and problem solving skills, which can be linked to social and economic characteristics both across and within countries.

The key objectives of the survey are to profile the distribution of prose literacy, document literacy, numeracy, analytic reasoning and health literacy in the adult population (aged 15 to 74 years), and to identify subpopulations whose performance in these skill domains may place them at risk.

The ALLS 2006 collected information between July 2006 and January 2007 from 8,988 private dwellings throughout non-remote areas of Australia. The sample design ensured that within each state and territory, each household had an equal chance of selection. Information was obtained from one person aged 15 to 74 years in the selected household. If there was more than one person of this age, the person interviewed was selected at random. The details on survey content, sampling and survey methods are available in the *Adult Literacy and Life Skills, Australia: User Guide* (ABS, 2006).

While the survey was initially developed by Statistics Canada, some minor adaptations to survey questions and exercises were made to suit the Australian context. The ALLS 2006 was conducted under the authority of the *Census and Statistics Act 1905*.

After completion of a background questionnaire, the randomly selected respondent completed a core task booklet (CTB). The CTB component is designed to identify respondents who are unlikely to be able to complete the exercises included in the main task booklet (MTB). The CTB contained six basic questions for the respondent to complete. Only respondents who correctly answered a minimum of three questions in the CTB moved on to the MTB. In all, 8,274 of 8,988 respondents completed the MTB (table 2.1).

In the ALLS 2006, each respondent was required to complete one booklet which consisted of tasks from two of the possible eight blocks of questions. This design, which is referred to as a Balanced Incomplete Block design, allows all the questions to be asked of a significant number of respondents, although not all the questions are asked of all the respondents in the survey.

The distribution of questions by literacy domain is shown in table 2.1. Since there were over 160 task lists, each booklet consisted of two (of a possible eight) blocks of questions with the total number of questions ranging from 17 to 53. The blocks of questions measure different skill domains: Blocks 1 to 4 measure prose and document literacy; Blocks 5 and 6 measure numeracy; and Blocks 7 and 8 measure problem solving. These blocks were then distributed across the 28 different booklets in differing combinations.

**2.1 Balanced Incomplete Block design, allocation of main task booklets, ALLS 2006**

| Booklet number | Blocks | | | | | | | | Number of questions | Respondents | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | | Number | Percent |
| 01 | X | X | | | | | | | 53 | 258 | 3.1 |
| 02 | X | | X | | | | | | 53 | 280 | 3.4 |
| 03 | | X | X | | | | | | 52 | 283 | 3.4 |
| 04 | | X | | X | | | | | 52 | 291 | 3.5 |
| 05 | X | | X | | | | | | 53 | 282 | 3.4 |
| 06 | | | X | X | | | | | 52 | 279 | 3.4 |
| 07 | X | | | X | | | | | 53 | 263 | 3.2 |
| 08 | | X | | X | | | | | 52 | 286 | 3.5 |
| 09 | X | | | | X | | | | 47 | 266 | 3.2 |
| 10 | | X | | | | X | | | 46 | 280 | 3.4 |
| 11 | | | X | | | X | | | 46 | 276 | 3.3 |
| 12 | | | | X | X | | | | 46 | 290 | 3.5 |
| 13 | | X | | | X | | | | 46 | 282 | 3.4 |
| 14 | | | X | | X | | | | 46 | 268 | 3.2 |
| 15 | X | | | | | X | | | 47 | 270 | 3.3 |
| 16 | | | | X | | X | | | 46 | 281 | 3.4 |
| 17 | | | | | X | X | | | 40 | 414 | 5.0 |
| 18 | | | | | X | X | | | 40 | 426 | 5.2 |
| 19 | X | | | | | | | X | 35 | 280 | 3.4 |
| 20 | | X | | | | | X | | 35 | 288 | 3.5 |
| 21 | | | X | | | | X | | 35 | 280 | 3.4 |
| 22 | | | | X | | | | X | 34 | 285 | 3.4 |
| 23 | X | | | | | | X | | 36 | 257 | 3.1 |
| 24 | | | | X | | | X | | 35 | 271 | 3.3 |
| 25 | | X | | | | | | X | 34 | 269 | 3.3 |
| 26 | | | X | | | | | X | 34 | 268 | 3.2 |
| 27 | | | | | | | X | X | 17 | 399 | 4.8 |
| 28 | | | | | | | X | X | 17 | 402 | 4.9 |
| Total Block repeats | 8 | 8 | 8 | 8 | 6 | 6 | 6 | 6 | 1,182 | 8,274 | 100.0 |
| # of main questions | 27 | 26 | 26 | 26 | 20 | 20 | 9 | 8 | 162 | | |
| Literacy dimension | PL/DL | PL/DL | PL/DL | PL/DL | NL | NL | PS | PS | | | |

Note: PL = Prose Literacy ; DL = Document Literacy ; NL = Numeracy ; PS = Problem solving

# 3. MEASUREMENT OF LITERACY

Although not every respondent answered all the questions for all the literacy domains, each respondent is given a score for each literacy domain, based upon their proficiency in their allocated MTB and responses to the background questionnaire. The detail on the how the scores for each literacy domain are computed is explained in Yammamoto (2002). For each of the four literacy domains, five plausible values are provided for each respondent. Since literacy scores are imputed for all individuals irrespective of which main task booklet they completed, or even if they did not complete a task booklet, users should take care when using these plausible scores as individual test scores. Yammamoto (2002:15) cautions users not to use population level literacy as a proxy for individual level test scores in that "… plausible values are not test scores for individuals in the usual sense". Since plausible values are constructed explicitly to provide consistent population estimates, these may not provide unbiased estimates of the literacy proficiency of individuals. Further, since these plausible values have been obtained by conditioning on respondents' background characteristics, any assessment of association of these values to the respondents' background characteristics might just reflect a spurious relationship (Carey *et al.*, 2000, p. 245). Carey *et al.* further argue for use of the more easily understood concept of the percentage of correct responses as an alternative for individual measurement. However, others have criticised the proportion of correct response as being a crude method of generating unbiased estimates of proficiency at the individual level (Boothby, 2005).

For each literacy domain, proficiency is measured on a scale ranging from 0 to 500 points. Each person's score denotes a point at which they have an 80 per cent chance of successfully completing tasks with a similar level of difficulty. To facilitate analysis, these continuous scores have been grouped into skill levels with Level 1 being the lowest measured level of literacy. The levels indicate specific sets of abilities, and therefore the thresholds for the levels are not equidistant. As a result, the ranges of scores in each level are not identical. In fact, for the prose literacy, document literacy, numeracy and health literacy domains, Level 1 captures almost half of the scale, from 0 to 225, for a range from 0 to 500. The thresholds for the problem solving domain are somewhat different, with Level 1 covering precisely half of the scale, from 0 to 250, for a range from 0 to 500.

In terms of the minimum level of literacy that is required to function in today's knowledge-based economy, Level 3 literacy is regarded as the threshold (Statistics Canada and OECD, 2005). In the analysis that follows, we have equated 'low literacy' with document literacy of Level 2 or below.

# 4. ESTIMATION METHOD

The method we have used to estimate literacy levels for small geographical areas is the random intercept logistic model. The dependent variable is the lower level of literacy indicator, where a value of 1 is assigned if the respondent falls into the 'low literacy' category, and 0 otherwise. As explained earlier, the two lowest literacy levels (Levels 1 and 2) were combined to create 'low literacy' category. As the dependent variable is binary, we use a logistic model. To capture area level variability, we use a random intercept logistic model. Our assumption is that individual literacy levels differ based on respondents' personal characteristics as well as where the individual resides. This assumes that respondents coming from the same area have similar literacy levels even after controlling for the effects of individual characteristics. In the event that literacy level is highly correlated within respondents from the same geographical area, fitting a model that ignores this clustering or grouping of individuals will lead to incorrect estimates.

We estimated the logistic coefficients with the STATA module `xtmelogit`, which fits mixed effects models for binary/binomial responses (StataCorp, 2007). Although `xtmelogit` allows users to specify the structure of covariance matrix for the random effects – independent, exchangeable, identity or unstructured – we specified the independent covariance structure, which is a default option. An independent covariance structure allows for a distinct variance for each random effect and assumes that all covariances are zero.

## 4.1 Variables

As explained already, the dependent variable is a low literacy level indicator, taking a value of one if the respondent's document literacy level is Level 2 or below, and zero if the document literacy level is Level 3 or above. The independent variables in the model have been selected on the basis of past studies on literacy (Willms, 2007; Reder, 1997), and also based on their availability in the 2006 Census of Population and Housing. The description and measurement of these variables are provided in table 4.1.

### 4.1 Description and measurement of model variables

| *Variable* | *Description* | *Census equivalent* |
|---|---|---|
| **Dependent variable** | | |
| Low literacy level | Document Literacy Level 1 or 2 | NA |
| **Independent variables** | | |
| Australian-born | Respondents reporting their country of birth as Australia. | Persons born in Australia (including external territories) |
| Male | Male respondents | Gender, male |
| Aged 15 to 24 years | Respondents aged between 15 and 24 years | Age in continuous years, and categorised in five year age groups. |
| Aged 25 to 44 years | Respondents aged between 25 and 44 years | Age in continuous years, and categorised in five year age groups. |
| Years of education: 8 years or below | Respondent with eight or fewer years of formal education | School education level: Year 8 or below |
| Years of education: 9–11 years | Respondent with 9–11 years of formal education | School education level: Year 9–11 |
| Years of education: 12 years | Respondent with 12 years of formal education | School education level: Year 12 |
| Years of education: 13–15 years | Respondent with 13–15 years of formal education | Highest completed non-school qualification: Certificate, Diploma or Advanced Diploma |
| Years of education: 16–17 years | Respondent with 16–17 years of formal education | Highest completed non-school qualification: Bachelor, Graduate certificate or Graduate Diploma |
| Years of education: 18 years or more | Respondent with 18 years or more of formal education | Highest completed non-school qualification: Postgraduate Degree |
| Employed | Respondent was employed during the survey week | Respondent employed full-time, part-time, or away from work |
| Poor English | Respondents considering that they speak English 'not well' or 'not at all' | Respondents considering that they speak English 'not well' or 'not at all' |
| Managerial or professional occupation | Respondents employed in a managerial or professional occupation (ISCO88 code between 1000 and 2500) | Respondents employed in a managerial or professional occupation (ISCO88 code between 1000 and 2500) |
| SEIFA IRSD decile | The decile of the SEIFA Index of Relative Socio-economic Disadvantage for LGAs as at 2006 Census | The decile of the SEIFA Index of Relative Socio-economic Disadvantage for LGAs as at 2006 Census |

Descriptive statistics based on estimates from ALLS 2006 are presented in table 4.2. Of the total responding sample, 46% were assessed as falling into the low literacy category (Document Literacy Level 1 or 2). These estimates have been obtained using unweighted data. The survey data show that more than one in four respondents were born overseas, while only 2 percent of all respondents reported that they have poor

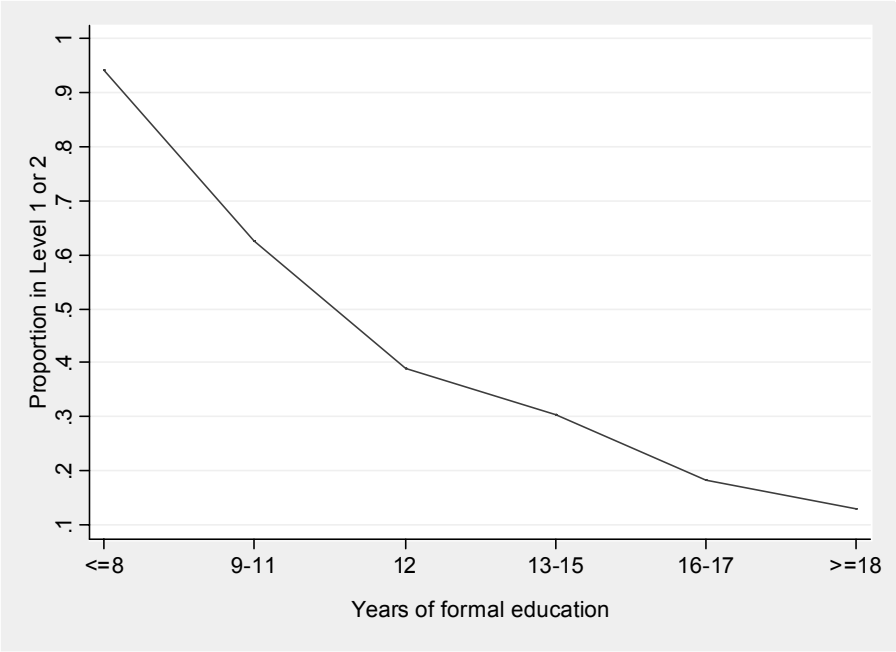
English speaking skills. Two-thirds of the respondents were employed during the survey. If we look at the socio-demographic characteristics of individuals in the low literacy category, we see that fewer of them were born in Australia, more of them were in older age category, few of them possessed tertiary qualification, or were employed, and of those who were employed proportionally more were working in non-managerial or non-professional jobs. Most importantly, nearly one in 20 reported that he/she had poor English language speaking skills.

**4.2 Descriptive statistics, ALLS 2006**

| | *Document Literacy Level* | | |
|---|---|---|---|
| | *Level 1–2* | *Level 3–5* | *Total* |
| Country of birth | | | |
| Born in Australia | 69.7% | 75.6% | 72.8% |
| Gender | | | |
| Male | 44.9% | 47.5% | 46.3% |
| Age group | | | |
| Aged 15 to 24 | 11.0% | 13.5% | 12.3% |
| Aged 25 to 44 | 28.6% | 48.3% | 39.2% |
| Aged 45 to 74 | 60.4% | 38.2% | 48.5% |
| Years of formal education | | | |
| 8 years or below | 18.3% | 1.0% | 9.0% |
| 9–11 years | 48.8% | 25.0% | 36.0% |
| 12 years | 12.9% | 17.4% | 15.3% |
| 13–15 years | 12.9% | 25.3% | 19.5% |
| 16–17 years | 4.8% | 18.3% | 12.0% |
| 18 years or higher | 2.3% | 13.1% | 8.1% |
| Employment status | | | |
| Employed | 52.9% | 80.9% | 67.9% |
| English speaking skills | | | |
| Speaks English 'not well' or 'not at all' | 4.5% | 0.4% | 2.3% |
| Occupation | | | |
| Manager or Professional | 8.8% | 27.9% | 19.1% |
| SEIFA decile | | | |
| Respondents from LGA in the lowest 3rd decile | 14.4% | 8.5% | 11.2% |
| Sample size | 4,160 | 4,828 | 8,988 |

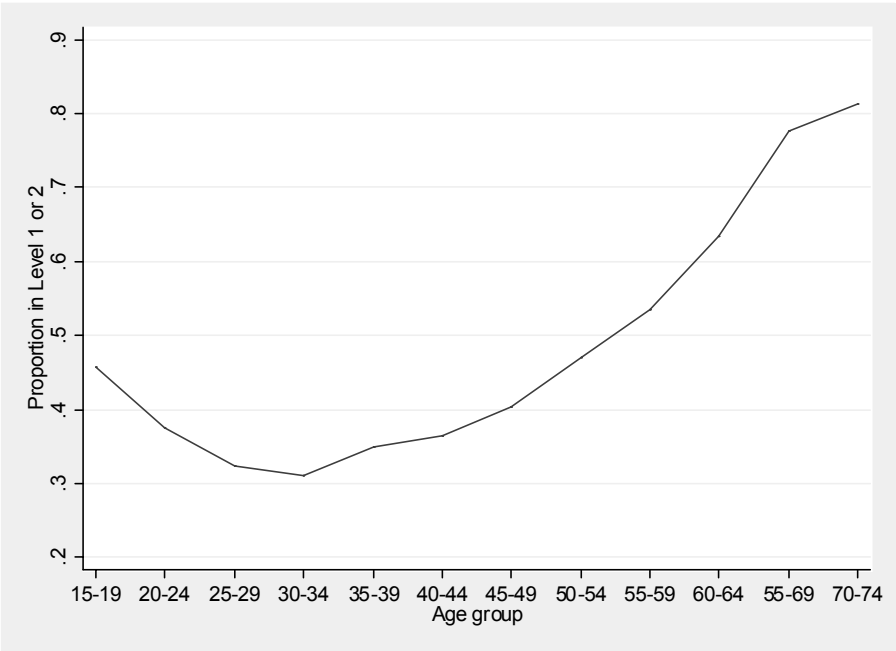
In figure 4.3 we present the observed prevalence of low literacy (i.e. in Levels 1 or 2) by number of years of formal education. The graph shows that as the years of formal education increase, the prevalence of low literacy decreases. For example, at eight or fewer years of total education, the prevalence of low literacy is nearly 100 percent. After 12 years of formal schooling, the mean prevalence reduces to just over 40 percent, and at 18 or more years of formal education, the prevalence of low literacy is less than 15 percent. The graph clearly shows the association between years of formal education and literacy level, although the association is not quite linear.

**4.3 Observed low literacy levels, by years of formal education, ALLS 2006**




In figure 4.4, we show the prevalence of low literacy by age of respondent (in five year age groups). The graph shows that, after the age of 30 years, the prevalence of low literacy increases with increasing age. Once again, the association between age and literacy is not linear. For the lowest age group (i.e. aged 15–19 years), the prevalence of low literacy is more than 45%. As age increases, the prevalence of low literacy decreases up to age 30–34 years, and then increases for the older age groups.

**4.4 Observed low literacy levels, by age of respondent, ALLS 2006**



We can also explore whether there is any association between a respondent's occupation and literacy. We speculate that managerial and professional occupations require higher levels of literacy than non-managerial and blue collar occupations. In figure 4.5 we show the prevalence of low literacy by years of formal education and occupation. The graph shows that, for the same number of years of formal education, persons employed as managers or professionals have a lower prevalence of low literacy than those employed in non managerial or non-professional occupations.

**4.5 Observed low literacy levels, by years of formal education and occupation, ALLS 2006**


## 4.2 Model

As explained already, in estimating low literacy levels, we fit a Bernoulli random intercept logistic regression model. In the model we include a Local Government Area-specific random intercept $u_j$ which is normally distributed with a mean 0 and variance equal to $\sigma_u^2$. The model is in the form:

$$\text{logit}\left\{\Pr\left(y_{ij}=1 \mid x_{ij}, u_{ij}\right)\right\} = \beta_0 + \beta_1 \text{ Australian-born}_{ij} + \beta_2 \text{ Male}_{ij}$$

$$+\beta_3 \text{ Aged } 15-24 \text{ years}_{ij} + \beta_4 \text{ Aged } 25-44 \text{ years}_{ij} + \beta_5 \text{ Tertiary education}_{ij}$$

$$+\beta_6 \text{ Post-secondary education}_{ij} + \beta_7 \text{ Secondary education}_{ij} + \beta_8 \text{ Employed}_{ij}$$

$$+\beta_9 \text{ Poor English}_{ij} + \beta_{10} \text{ Manager or professional}_{ij} + \beta_{11} \text{ SEIFA decile}_{j} + u_j$$

where $i$ = person, $j$ = small area (LGA) and $u_j \sim N(0, \sigma_u^2)$ is an area level effect.

## 4.3 Results from the random intercept logistic model

The odds ratios and associated standard errors obtained from the random intercept logistic regression model are shown in table 4.6. Detailed results are given in table A.2 in the Appendix. The results show that, as expected, the higher the level of education, the lower the odds of falling into the low literacy category. By contrast, with increased age, the odds of falling in the low literacy category increases. Country of birth also has a significant relationship with low literacy. Australian-born respondents have lower odds of falling into the low literacy category compared to overseas-born respondents (0.7 to 1). Those in employment and those working in managerial or professional occupations also have lower odds of low literacy. Additionally, people who report that they speak English 'not well' or 'not at all' have the highest odds (6.8 to 1) of falling into the low literacy category. Although gender is assumed to be predictor of literacy, we found no significant gender difference in literacy levels, once the effects of other variables were controlled for.

**4.6 Predicting prevalence of low literacy with the random intercept logistic regression, ALLS 2006**

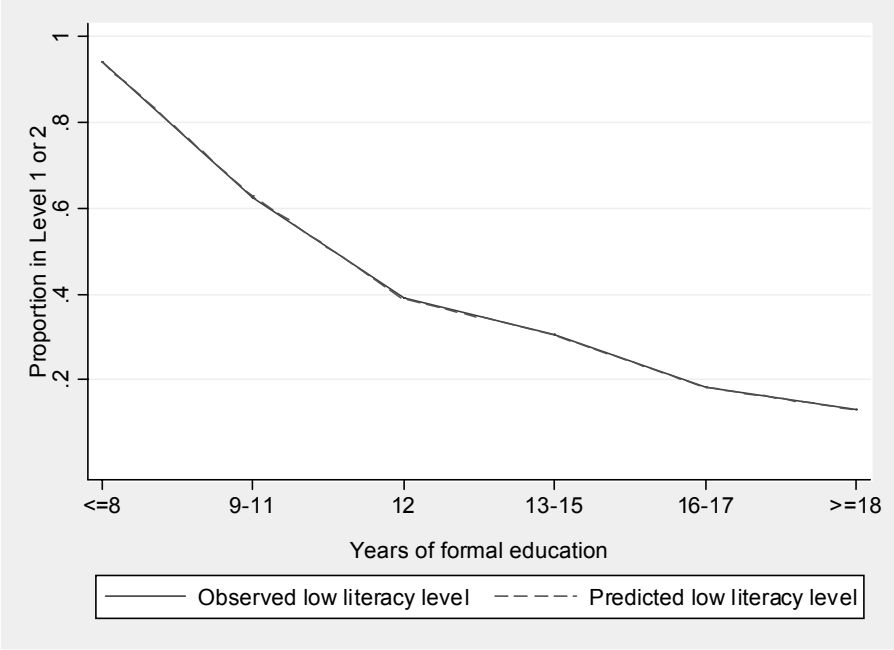
| *Variable* | *Odds ratio* | *Standard error* | *z* |
|---|---|---|---|
| Country of birth (Base = Overseas-born) | | | |
| Australian-born | 0.70 | 0.042 | –6.0 |
| Gender (Base = Female) | | | |
| Male | 1.00 | 0.051 | –0.1 |
| Age group (Base = Aged 45 to 74 years) | | | |
| Aged 15 to 24 years | 0.58 | 0.046 | –6.9 |
| Aged 25 to 44 years | 0.64 | 0.036 | –8.0 |
| Years of formal education (Base = 8 years or fewer) | | | |
| 9–11 years | 0.15 | 0.025 | –11.7 |
| 12 years | 0.07 | 0.011 | –16.2 |
| 13–15 years | 0.05 | 0.008 | –18.3 |
| 16–17 years | 0.03 | 0.005 | –20.3 |
| 18 years or higher | 0.02 | 0.003 | –20.8 |
| Employment status (Base = Other) | | | |
| Employed | 0.53 | 0.031 | –11.0 |
| English speaking skills (Base = Speaks English 'well' or 'very well') | | | |
| Speaks English 'not well' or 'not at all' | 6.81 | 1.907 | 6.9 |
| Occupation (Base = Other occupation) | | | |
| Manager or Professional | 0.53 | 0.039 | –8.6 |
| SEIFA decile (Base = fourth decile or higher) | | | |
| LGA in the lowest three deciles | 1.31 | 0.124 | 2.8 |
| Random effects parameter | | | |
| LGA level variance | 0.07 | 0.023 | |

The SEIFA score, a contextual variable measuring area level disadvantage, is also a significant predictor of low literacy. This suggests that personal as well contextual level characteristics impact on literacy at the individual level. The results also show that, even after controlling for the individual and contextual characteristics, there is significant LGA level variability in literacy. In other words, a significant portion of residual variability in literacy that is not explained by individual or contextual characteristics can be explained by the area individuals live in. If we had not used the random intercept model, we would not have captured this area level variability.

In figure 4.3 we plotted the observed prevalence of low literacy by years of formal education. To assess how well the model fits the data, we can plot the predicted prevalence of low literacy against years of formal education.

In figure 4.7 we have plotted the observed and predicted low literacy rate by years of formal education. To obtain the predicted prevalence of low literacy, we added all the predicted probabilities for sampled respondents and estimated means for each education level.

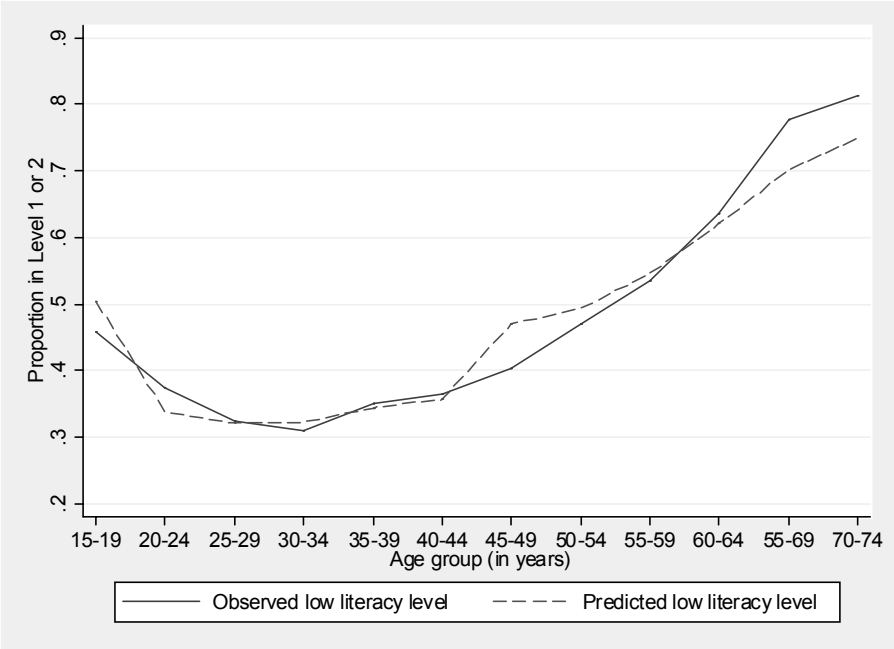
**4.7 Observed and predicted low literacy levels, by years of formal education, ALLS 2006**




The graph shows that there is a very close agreement between the observed and predicted low literacy, by years of formal schooling. The two lines are on top of each other. This suggests that model is fitting the data well.

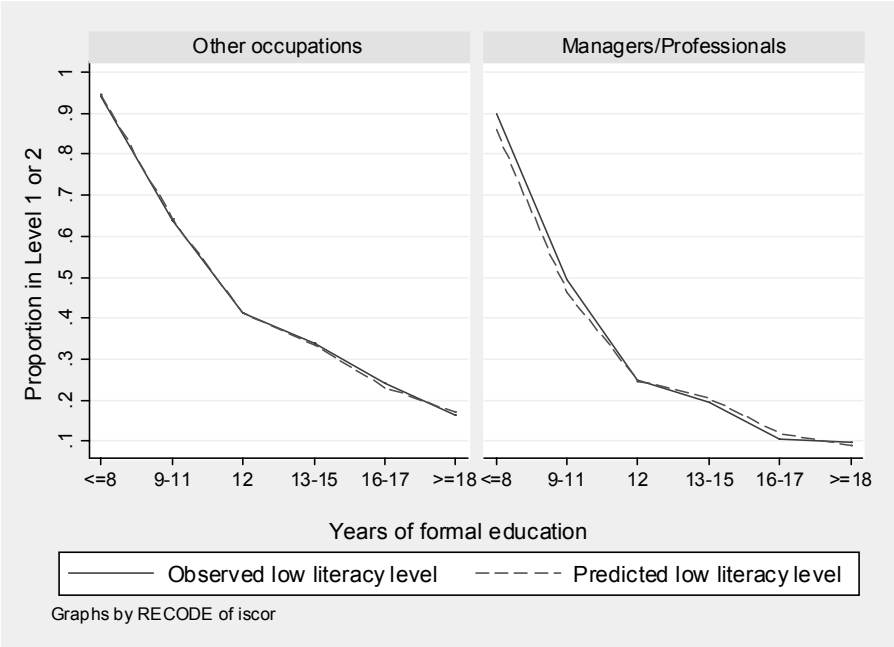
Similarly, in figure 4.8 we show the observed and predicted low literacy by age of respondent. The predicted low literacy level is very similar to the observed level, indicating that again the model fits the data well.

### 4.8 Observed and predicted low literacy levels, by age, ALLS 2006



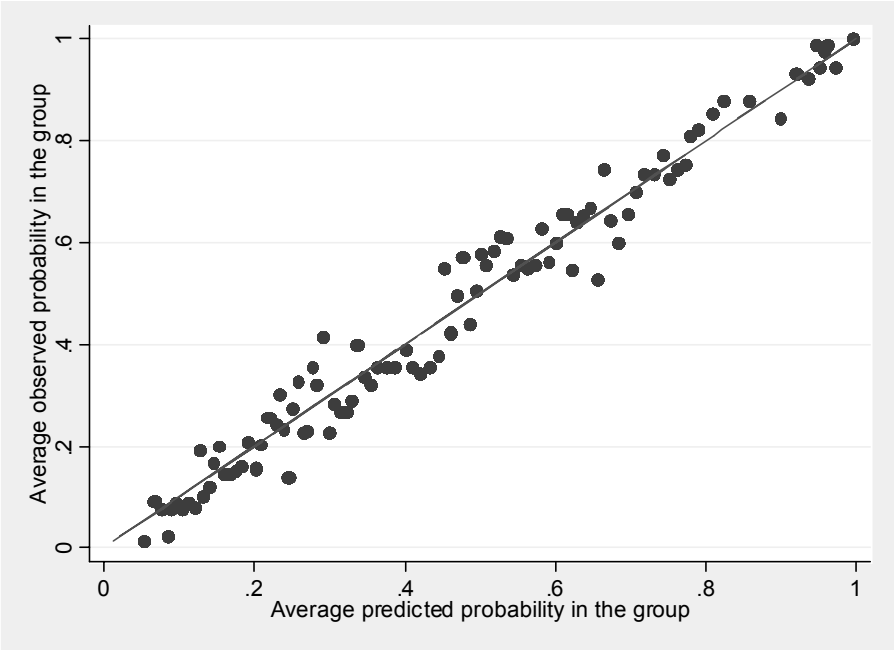

Finally, we looked at the predicted and observed low literacy by occupation for different education levels, and found that the predicted prevalence of low literacy compares well with observed levels, except among poorly educated managers and professionals.

### 4.9 Observed and predicted low literacy levels, by education and occupation, ALLS 2006



Another diagnostic of model fit is to compare the predicted and observed probabilities of low literacy within a propensity group (Pfeffermann, Terryn and Moura, 2005). This involves grouping observations into propensity groups based on estimated probabilities of low literacy. We divided respondents into 100 groups in order of predicted probability of low literacy. Then we calculated the average observed prevalence of low literacy and average predicted probability of low literacy, and plotted them. If all the points lie in a straight line, this tends to signify a good fit of the model. In our case, the observed and predicted probabilities show a linear trend.

**4.10 Observed and predicted probabilities of low literacy, by propensity groups, logistic model with random effects, ALLS 2006**




When a Pearson chi-square test was conducted on the observed and expected probabilities on these propensity groups, we found an insignificant chi-squared value of 0.68 (df=99).

$$\text{Pearson residual} = \frac{o_k - e_k}{\sqrt{e_k}}$$

where

$o_k$ = observed proportion in propensity group $k$;

$e_k$ = predicted proportion in propensity group $k$;

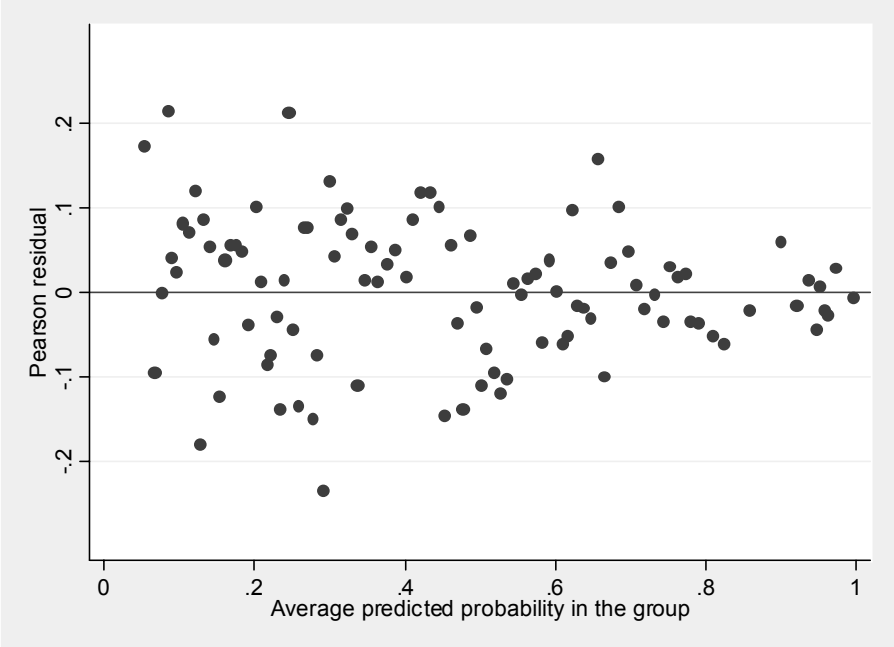
$k$ = group;

and

$$\sum_k \frac{(o_k - e_k)^2}{e_k} \sim \chi^2$$

The plot of the Pearson residuals against predicted probability for these propensity groups (figure 4.11) shows that the residuals are generally unbiased and small in magnitude. However, there is some indication that propensity groups with a low prevalence of low literacy have slightly larger residuals.

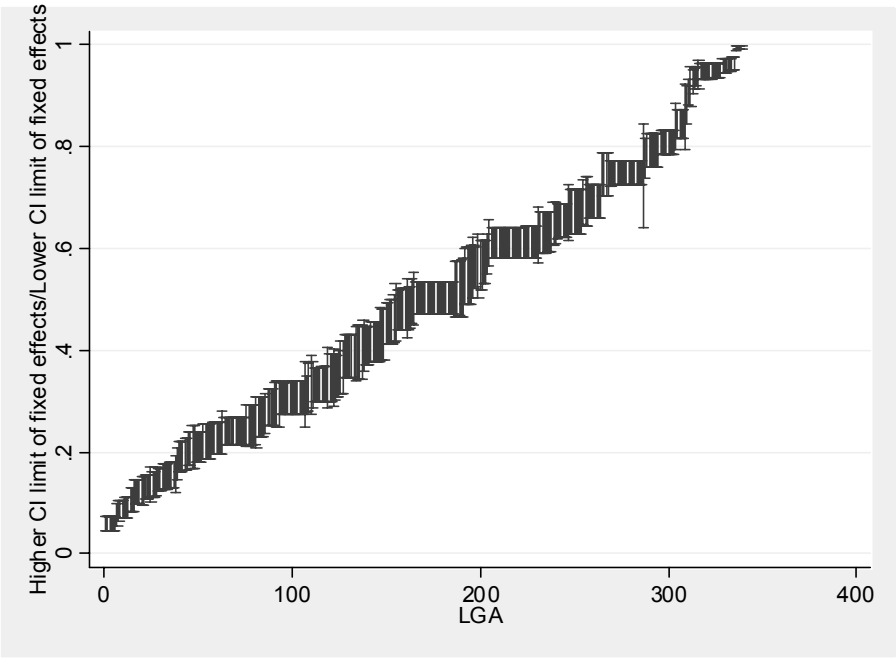
**4.11 Plot of Pearson residuals and predicted probability, by propensity group**



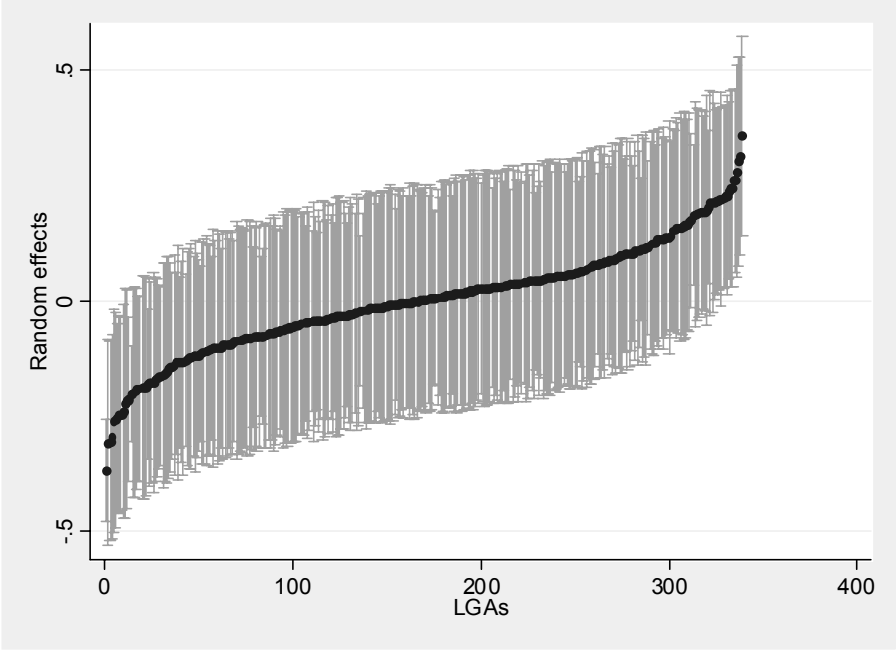

The evidence so far suggests that the random intercept model for predicting low literacy at individual level is reasonably good, in that the model predicts the observed literacy level fairly well. In the next section we describe how we used the logistic coefficients obtained from unit level model to estimate prevalence of low literacy at the Local Government Area level.

In order to see how areas vary in terms of literacy levels, we have plotted the fixed effects for each area (figure 4.12) and area level effects (figure 4.13) and their confidence intervals. The results show that (a) there is a high variability in literacy level across LGAs, and (b) there is a wide range of area level effects on literacy, even after controlling for individual and contextual characteristics.

### 4.12 Prevalence of low literacy at the LGA level – Fixed effects only




### 4.13 Prevalence of low literacy at the LGA level – Area level effect



## 5. PREDICTION METHOD

Local Government Areas (LGAs) have been treated as small areas for the purposes of this study. The purpose of the study is to find $p_j$, the proportion of adults aged 15 to 74 years in the $j$th local area who are in literacy levels 1 and 2.

$$p_j = \frac{\sum y_{ij}}{N_j}$$

where $N_j$ is the population of adults in area $j$, and $y_{ij}$ indicates whether or not the individual falls into the low literacy category. Since we want to estimate the proportion $p_j$, we do so by (Royall, 1970):

$$\hat{p}_j = \frac{\left( \sum_{i \in S} y_{ij} + \sum_{i \notin S} \hat{\pi}_{ij} \right)}{N_j}$$

where

$\Sigma y_{ij}$ = sum of the values of the low literacy indicator for sampled individuals from the $j$th Local Government Area, and

$\Sigma \hat{\pi}_{ij}$ = sum of the estimated probabilities for non-sampled individuals in the $j$th Local Government Area.

To obtain $\hat{\pi}_{ij}$, we employ the model-based approach proposed by Dempster and Tomberlin (1980). Under this approach, a model that describes the probabilities associated with individuals in the population is:

$$y_{ij} \mid \pi_{ij} \sim \text{ i.i.d. Bernoulli } (\pi_{ij}) \; ; \; \text{logit } (\pi_{ij}) = X^T \beta + u_j$$

At the local government area level, the prediction is obtained by utilising parameters obtained from person level model:

$$\text{logit } (\pi_j) = X^{\mathrm{T}} \beta + u_j$$

where

$X^{\mathrm{T}}$ = a vector of predictor variables associated with fixed effects at the area level;

$\beta$ = a vector of fixed effects logistic regression parameters obtained from person level model; and

$u_j$ = random effects, estimated from person level model.

The vector of predictor variables is expressed as the percentage of total population aged 15 to 74 years in the area with certain attributes. For example, the variable measuring country of birth, 'Australian-born', at a small area is obtained by counting all the people who were born in Australia in that small area divided by total population in that area aged 15 to 74 years.

Once $\text{logit}(\pi_j)$ is estimated, $\hat{\pi}_j$ is calculated by exponentiation as follows:

$$\hat{\pi}_j = \frac{\exp\left(X^{\mathrm{T}}\hat{\beta} + \hat{u}_j\right)}{1 + \exp\left(X^{\mathrm{T}}\hat{\beta} + \hat{u}_j\right)}$$

For those small areas where no sample was selected, the estimated probability is equal to the synthetic estimates, with random effects set to zero.

## 5.1 Variance estimation

Saei and Chambers (2003, p. 27) have shown that the EBLUP mean square error (correctly known as the Mean Cross Product Error or MCPE) can be estimated by:

$$\widehat{\text{MCPE}}\left(\hat{\theta}\right) = \hat{\mathbf{G}}_1 + \hat{\mathbf{G}}_2 + 2\hat{\mathbf{G}}_3 + \hat{\mathbf{G}}_4$$

where

$$\hat{\mathbf{G}}_1 = \mathbf{Z}_r^{+}\,\hat{\mathbf{T}}_s^{*}\,\mathbf{Z}_r^{+\prime}$$

$$\hat{\mathbf{G}}_2 = \left[\mathbf{X}_r^{+} - \mathbf{Z}_r^{+}\,\hat{\mathbf{T}}_s^{*}\,\mathbf{Z}'\,\hat{\mathbf{B}}_s\mathbf{X}\right]\hat{\mathbf{T}}_{11}\left[\mathbf{X}_r^{+\prime} - \mathbf{X}'\,\hat{\mathbf{B}}_s'\,\mathbf{Z}_r^{+}\,\hat{\mathbf{T}}_s^{*}\,\mathbf{Z}_r^{+\prime}\right]$$

$$\hat{\mathbf{G}}_3 = \left[\text{tr}\left(\hat{\nabla}_\alpha\,\boldsymbol{\Sigma}_s^{+}\,\hat{\nabla}'_{\alpha'}\right)\widehat{\text{Var}}\left(\hat{\varphi}\right)\right]$$

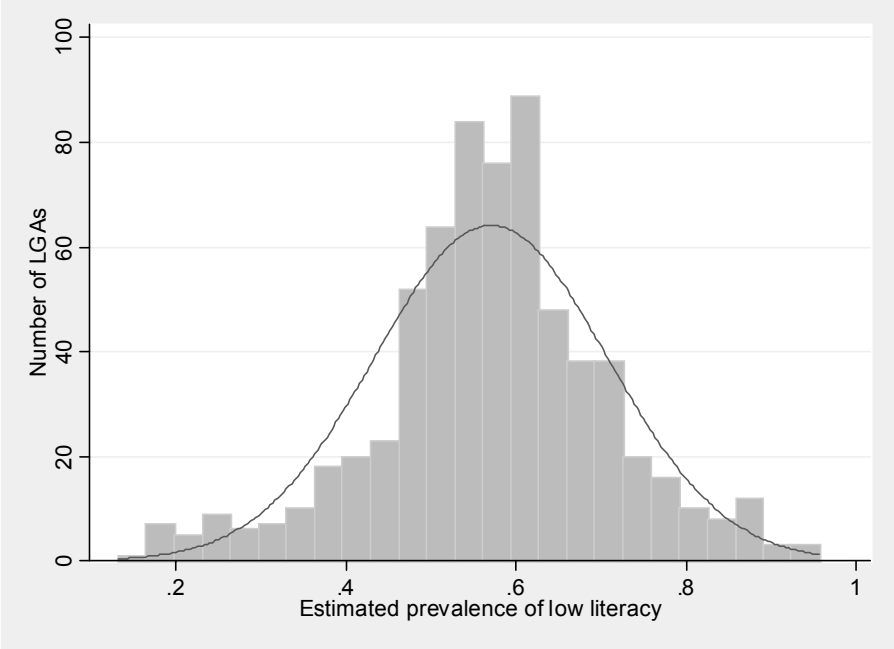
$$\hat{\mathbf{G}}_4 = \mathbf{a}_r\,\hat{\mathbf{B}}_r\,\mathbf{a}'_r$$

Further details about the methodology for estimating the MCPE can be found in Saei and Chambers (2003). We are currently working to implement this method.

## 5.2 Predicted results

The distribution of the estimated prevalence of low literacy level at LGA is shown in figure 5.1. The distribution shows that the prevalence of low literacy is leptokurtic (a kurtosis of 3.7) with higher peaks around the mean compared to the normal distribution. These peaks result from the data being highly concentrated around the mean, due to lower variation within LGAs. The figure shows that the majority of LGAs have the predicted prevalence of low literacy close to the mean, and few LGAs have either very high or very low prevalence of predicted low literacy.

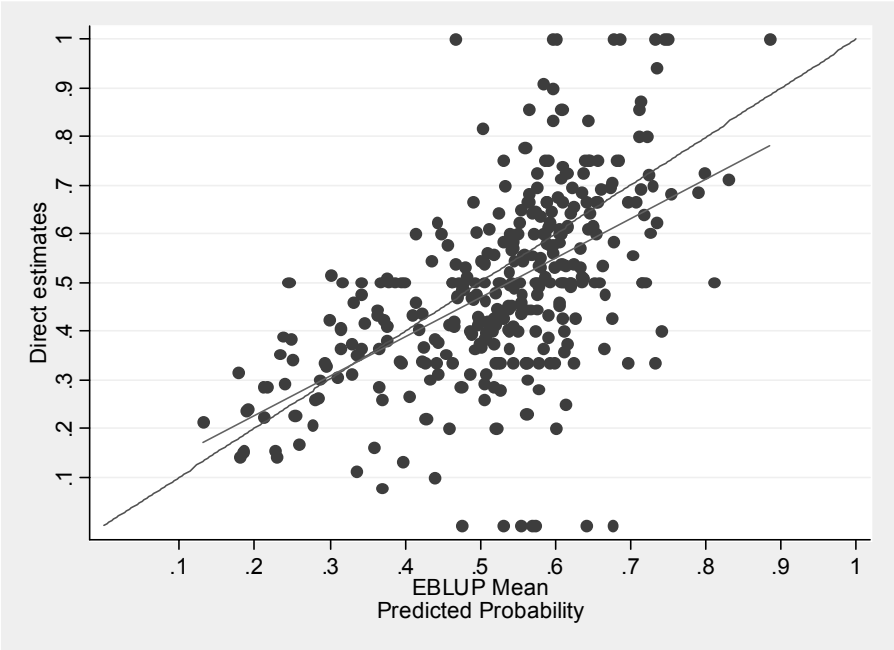
**5.1 Histogram of estimated prevalence of low literacy at the LGA level**




When we compare the direct estimates and modelled estimates of the prevalence of low literacy at the LGA level, we find that higher direct estimates correspond to higher predicted estimates (figure 5.2). However, the direct estimates of the prevalence of low literacy range from 0 to 100%, while the predicted estimates range from 13 to 89% only. This means that there is shrinkage in estimation and that very high or very low direct estimates cannot be predicted accurately by the estimation method. When the sample size for a LGA is too small, it is very likely to have direct estimates at the extreme end. For example, of the 339 LGAs sampled, 13 LGAs had two or fewer respondents in the survey, and a further 29 LGAs with five or fewer respondents (see table A.1 in the Appendix). Extreme direct estimates of low literacy at the LGA level are inevitable when the sample size as small as this.

Although auxiliary variables in the Census have the same definition as that of the survey variables, we suspect that the measurement could be different due to mode effects. This is an issue we are investigating further.

**5.2 Scatter plot, direct estimates and EBLUP, prevalence of low literacy in sampled LGAs**




## 5.3 Comparison with other measures of disadvantage

To explore how estimates of the prevalence of low literacy compare with other measures of disadvantage at the small area level, we examined the socio-economic status of the 20 LGAs having the highest prevalence estimates and the 20 LGAs having the lowest prevalence estimates. The ABS produces area level disadvantage measures for various geographical levels. We used the 2006 Index of Relative Socio-economic Disadvantage (IRSD). Most of the variables that are used in the creation of the IRSD index relate to the age, country of birth, education, employment and occupation profile of the area. These are the same variables we have used in our estimation and prediction of low literacy at the LGA level.

In table 5.3 we list the 20 LGAs having the highest estimated prevalence of low literacy. All but one of these LGAs were not included in the ALLS 2006 sample, which deliberately excluded rural and remote areas. Almost all LGAs listed in table 5.3 are from the Northern Territory, and are small in terms of population size. The model predicted the prevalence of low literacy in these LGAs to lie within the range 0.85–0.96. These LGAs have also been categorised as highly disadvantaged areas, falling into the lowest decile of the IRSD index.

The data in table 5.3 indicate that, although the prevalence of low literacy may not have been predicted accurately, the prevalence of low literacy at the small area level can be used as another measure of disadvantage. We are aware that purely synthetic small area predictions for out-of-sample areas have known problems. These synthetic estimates are highly dependent upon how good the unit level model is, and whether the strength of association obtained from the unit level model is portable at the area level.

**5.3 List of Local Government Areas with the highest levels of low literacy**

| *LGA name* | *State* | *Population aged 15 to 74* | *Sample size* | *IRSD decile* | *Observed low literacy* | *Predicted low literacy* |
|---|---|---|---|---|---|---|
| Marngarr (CGC) | NT | 202 | 0 | 1 | NA | 0.96 |
| Thamarrurr (CGC) | NT | 1,195 | 0 | 1 | NA | 0.95 |
| Jilkminggan (CGC) | NT | 156 | 0 | 1 | NA | 0.93 |
| Alpurrurulam (CGC) | NT | 278 | 0 | 1 | NA | 0.92 |
| Lajamanu (CGC) | NT | 409 | 0 | 1 | NA | 0.92 |
| Numbulwar Numburindi (CGC) | NT | 446 | 0 | 1 | NA | 0.90 |
| Arltarlpilta (CGC) | NT | 155 | 0 | 1 | NA | 0.89 |
| Nyirranggulung Mardrulk Ngadberre (CGC) | NT | 828 | 7 | 1 | 1.00 | 0.89 |
| Umagico (S) | QLD | 131 | 0 | 1 | NA | 0.88 |
| Injinoo (S) | QLD | 252 | 0 | 1 | NA | 0.88 |
| Anmatjere (CGC) | NT | 635 | 0 | 1 | NA | 0.88 |
| Yuendumu (CGC) | NT | 501 | 0 | 1 | NA | 0.88 |
| Kunbarllanjnja (CGC) | NT | 599 | 0 | 1 | NA | 0.88 |
| Walangeri Ngumpinku (CGC) | NT | 247 | 0 | 1 | NA | 0.87 |
| Yugul Mangi (CGC) | NT | 1,046 | 0 | 1 | NA | 0.87 |
| Timber Creek (CGC) | NT | 337 | 0 | 1 | NA | 0.87 |
| Mer (IC) | QLD | 295 | 0 | 1 | NA | 0.86 |
| Elliott District (CGC) | NT | 292 | 0 | 1 | NA | 0.86 |
| Ltyentye Purte (CGC) | NT | 351 | 0 | 1 | NA | 0.85 |
| Badu (IC) | QLD | 499 | 0 | 1 | NA | 0.85 |

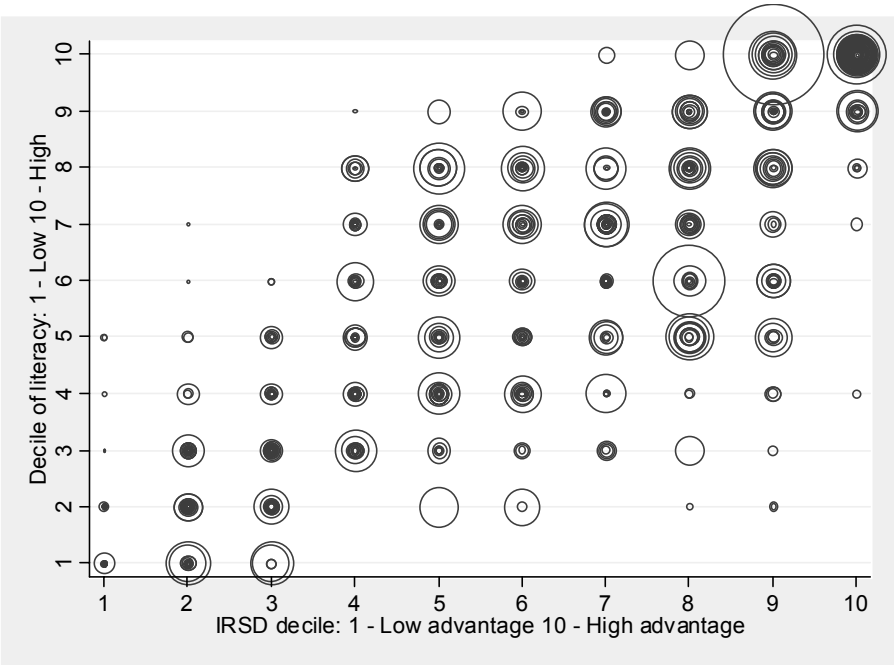
When we listed the 20 LGAs with the lowest estimated prevalence of low literacy, we first noticed that most were in the ALLS 2006 sample (18 out of 20 LGAs), and most were large in population terms (table 5.4). The direct estimates of proportions of low literacy in these LGAs ranged from a low of 0.14 to a high of 0.50. The corresponding range for predicted prevalence ranged from a low of 0.13 to a high of 0.25. All LGAs were found to lie in the tenth IRSD decile. Once again, the areas that were predicted to have low levels of low literacy were identified by the IRSD measure to be among the least disadvantaged. The results indicate that the small area estimates of literacy appear to be consistent with the IRSD measure.

**5.4 List of Local Government Areas with the lowest levels of low literacy**

| *LGA name* | *State* | *Population aged 15 to 74* | *Sample size* | *IRSD decile* | *Observed low literacy* | *Predicted low literacy* |
|---|---|---|---|---|---|---|
| Ku-ring-gai (A) | NSW | 70,430 | 33 | 10 | 0.21 | 0.13 |
| Mosman (A) | NSW | 19,339 | 0 | 10 | NA | 0.17 |
| Cambridge (T) | WA | 16,827 | 38 | 10 | 0.32 | 0.18 |
| Claremont (T) | WA | 6,524 | 14 | 10 | 0.14 | 0.18 |
| Nedlands (C) | WA | 14,926 | 20 | 10 | 0.15 | 0.18 |
| Woollahra (A) | NSW | 38,763 | 13 | 10 | 0.15 | 0.19 |
| Lane Cove (A) | NSW | 22,792 | 21 | 10 | 0.24 | 0.19 |
| Boroondara (C) | VIC | 112,957 | 54 | 10 | 0.24 | 0.19 |
| Unincorporated ACT | ACT | 250,327 | 430 | 10 | 0.29 | 0.21 |
| Willoughby (C) | NSW | 48,948 | 18 | 10 | 0.22 | 0.21 |
| North Sydney (A) | NSW | 50,502 | 35 | 10 | 0.29 | 0.22 |
| Manly (A) | NSW | 27,996 | 13 | 10 | 0.15 | 0.23 |
| Peppermint Grove (S) | WA | 1,138 | 7 | 10 | 0.14 | 0.23 |
| Cottesloe (T) | WA | 5,469 | 0 | 10 | NA | 0.23 |
| Hornsby (A) | NSW | 109,930 | 17 | 10 | 0.35 | 0.23 |
| Subiaco (C) | WA | 13,838 | 18 | 10 | 0.39 | 0.24 |
| Burnside (C) | SA | 30,434 | 41 | 10 | 0.29 | 0.24 |
| Hunter's Hill (A) | NSW | 9,231 | 2 | 10 | 0.50 | 0.24 |
| Leichhardt (A) | NSW | 39,582 | 26 | 10 | 0.38 | 0.25 |
| Stonnington (C) | VIC | 71,737 | 35 | 10 | 0.34 | 0.25 |

If we divide Local Government Areas into deciles based on predicted low literacy and compare these against IRSD deciles, we see that LGAs that have a high proportion of low literate adults also tend to be relatively more disadvantaged (figure 5.5). In the bubble plots, there are 676 bubbles each representing one LGA. The area of symbol is proportional to Australia's population aged 15–74 years. We see that large LGAs are relatively more advantaged than the smaller ones. It is equally possible that since the small LGAs are less likely to be included in the sample, the estimates for these out-of sample areas are purely synthetic and could be biased.

**5.5 Bubble plot, IRSD decile and low literacy decile**



## 6. CONCLUSION

The aim of this paper was to explore a method to generate experimental estimates of adult literacy for Local Government Areas (LGAs). We have shown that it is possible to generate estimates of the prevalence of low literacy for small areas, even though these areas may not have enough sample or may have no sample at all. When areas that were found to have a high prevalence of low literacy were compared with the ABS disadvantage measure (IRSD), we noticed that these areas were among the most disadvantaged areas in Australia. Similarly, areas where the prevalence of low literacy was low were found to be among the least disadvantaged on the IRSD measure. Although it is feasible to estimate the prevalence of low literacy within small areas, care must be exercised when using these measures. We have not yet tested the quality of these estimates and, as with any survey estimates, there is uncertainty attached to these estimates. Our next step is to calculate the measurement error surrounding these estimates and assess whether they can be of relevance for policy formulation.

## ACKNOWLEDGEMENTS

The author acknowledges Daniel Elazar, Chris Duncan, Caroline Daley and Peter Rossiter for their constructive comments, suggestions and technical help.

## REFERENCES

Australian Bureau of Statistics (2006) *Adult Literacy and Life Skills, Australia: User Guide*, cat. no. 4228.0.35.002, ABS, Canberra.

Boothby, D. (2005) *International Adult Literacy Survey: Literacy Skills, Occupational Assignment and Returns to Over- and Under-Education*, catalogue no. 89-552-MIE, no. 9, Statistics Canada.

Carey, S.; Bridgwood, A. and Thomas, M. (2000) *Measuring Adult Literacy: The International Adult Literacy Survey in the European Context*, Office for National Statistics, London.

Dempster, A.P. and Tomberlin, T.J. (1980) "The Analysis of Census Undercount from a Post-Enumeration Survey", *Proceedings of the Conference on Census Undercount*, Arlington, Virginia, pp. 88-94.

Organisation for Economic Cooperation and Development (1992) *Adult Literacy and Economic Performance*, OECD, Paris.

Organisation for Economic Cooperation and Development and Statistics Canada (1995) *Literacy, Economy and Society: Results of the First International Adult Literacy Survey*, OECD, Paris.

Pfeffermann, D.; Terryn, B. and Moura, F. (2005) *Small Area Estimation under a Two Part Random Effects Model. Application to Estimation of Literacy in Developing Countries*, International Conference on Survey Research Methods: Maximising Data Value, Data Use & Re-Use, organised by the Association for Survey Computing, the Office for National Statistics, the Market Research Society and the Royal Statistical Society, Newland Park, 15–16 September 2005.

Reder, S. (1997) *Synthetic Estimates of Literacy Proficiency for Small Census Areas*, Division of Adult Education and Literacy, Office of Vocational and Adult Education, US Department of Education.

Royall, R.M. (1970) "On Finite Population Sampling Theory under Certain Linear Regression Models", *Biometrika*, 57, pp. 377–387.

Saei, A. and Chambers, R. (2003) *Small Area Estimation Under Linear and Generalized Linear Mixed Models With Time and Area Effects*, S3RI Methodology Working Papers, M03/15, Southampton Statistical Sciences Research Institute, Southampton.
Available at <http://eprints.soton.ac.uk/8165/>

StataCorp (2007) *Stata Statistical Software: Release 10*, StataCorp LP, College Station, Texas.

Statistics Canada and Organisation for Economic Cooperation and Development (2005) *Learning a Living: First Results from the Adult Literacy and Life Skills Survey*, OECD, Paris.

Willms, J.D. (2007) *The Geographical Distribution of Adult Literacy Skills in Canada*, Human Resources and Social Development Canada, catalogue no. HS28-118/2007-MRC.

Yammamoto, K. (2002) *Estimating PISA students on the IALS Prose Literacy Scale*, Educational Testing Service, Princeton, New Jersey.

# APPENDIX

## A.1 Sample size distribution

| Number of respondents | Number of LGAs |
|---|---|
| 0 | 337 |
| 1 | 6 |
| 2 | 7 |
| 3–5 | 29 |
| 6–10 | 77 |
| 11–15 | 51 |
| 16–20 | 27 |
| 21–50 | 99 |
| 51–100 | 35 |
| > 100 | 8 |
| Total | 676 |

## A.2 Mixed-effects logistic regression, predicting low literacy at LGA level

| *Variable* | *Coefficient* | *Standard error* | *z* |
|---|---|---|---|
| Country of birth (Base = Overseas-born) | | | |
| Australian-born | –0.36 | 0.060 | –6.0 |
| Gender (Base = Female) | | | |
| Male | 0.00 | 0.051 | –0.1 |
| Age group (Base = Aged 45 to 74 years) | | | |
| Aged 15 to 24 years | –0.54 | 0.078 | –6.9 |
| Aged 25 to 44 years | –0.45 | 0.056 | –8.0 |
| Years of formal education (Base = 8 years or fewer) | | | |
| 9–11 years | –1.87 | 0.160 | –11.7 |
| 12 years | –2.70 | 0.167 | –16.2 |
| 13–15 years | –3.03 | 0.166 | –18.3 |
| 16–17 years | –3.61 | 0.178 | –20.3 |
| 18 years or higher | –4.04 | 0.195 | –20.8 |
| Employment status (Base = Other) | | | |
| Employed | –0.63 | 0.057 | –11.0 |
| English speaking skills (Base = Speaks English "well" or "very well") | | | |
| Speaks English "not well" or "not at all" | 1.92 | 0.280 | 6.9 |
| Occupation (Base = Other occupation) | | | |
| Manager or Professional | –0.64 | 0.074 | –8.6 |
| SEIFA decile (Base = fourth decile or higher) | | | |
| LGA in the lowest three deciles | 0.27 | 0.095 | 2.8 |
| Random-effects parameter | | | |
| LGA level variance | 0.07 | 0.023 | |

LR test vs. logistic regression: chibar2(01) = 23.90, Prob>=chibar2 = 0.0000

The Likelihood Ratio test (LR test) assesses whether a random-intercept model (as listed) is different to a simple logistic regression. Since the p-value is close to zero, we accept that the random coefficient model is different and that the variance component or area level variance should be taken into consideration.

Wald $\chi^2(13) = 1647.38$

This is the Wald chi-squared statistic. It is used to test the hypothesis that at least one of the coefficients is not equal to zero. The number in the parentheses indicates the degrees of freedom of the chi-squared distribution used to test the Wald chi-squared statistic and is defined by the number of predictors in the model (13).

Log-likelihood = –4824.2382, Prob > $\chi^2$ = 0.0000

This is the log-likelihood of the fitted model. It is used in the Likelihood Ratio chi-squared test of whether all predictors' coefficients in the model are simultaneously zero. As the p-value is close to zero, we conclude that all the coefficients are not simultaneously zero.

## FOR MORE INFORMATION . . .

*INTERNET* **www.abs.gov.au** the ABS website is the best place for data from our publications and information about the ABS.

### INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

*PHONE* 1300 135 070

*EMAIL* client.services@abs.gov.au

*FAX* 1300 135 211

*POST* Client Services, ABS, GPO Box 796, Sydney NSW 2001

## FREE ACCESS TO STATISTICS

All statistics on the ABS website can be downloaded free of charge.

*WEB ADDRESS* www.abs.gov.au