Catalogue no. 89-552-MIE — No.18
ISSN: 1480-9516
ISBN: 978-0-662-47311-4

## International Adult Literacy Survey

# Literacy and the Labour Market:

## The Generation of Literacy and Its Impact on Earnings for Native Born Canadians

by David A. Green and W. Craig Riddell

Culture, Tourism and the Centre for Education Statistics Division
Main Building, Room 2001, Ottawa, K1A 0T6

Telephone: 1-800-307-3382 Fax: 1-613-951-9040

Statistics Canada Statistique Canada

Canada

## How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website at www.statcan.ca or contact us by e-mail at infostats@statcan.ca or by phone from 8:30am to 4:30pm Monday to Friday at:

| | |
|---|---|
| **Toll-free telephone** (Canada and the United States): | |
| Enquiries line | 1-800-263-1136 |
| National telecommunications device for the hearing impaired | 1-800-363-7629 |
| Fax line | 1-877-287-4369 |
| Depository Services Program enquiries line | 1-800-635-7943 |
| Depository Services Program fax line | 1-800-565-7757 |
| | |
| **Statistics Canada national contact centre:** | 1-613-951-8116 |
| Fax line | 1-613-951-0581 |

## Information to access the product

This product, catalogue no. 81-552-MIE, is available for free in electronic format. To obtain a single issue, visit our website at www.statcan.ca and select Publications.

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1-800-263-1136. The service standards are also published on www.statcan.ca under About us > Providing services to Canadians.

# International Adult Literacy Survey

# Literacy and the Labour Market:

## The Generation of Literacy and Its Impact on Earnings for Native Born Canadians

**by David A. Green and W. Craig Riddell**
Department of Economics, University of British Columbia

Published by authority of the Minister responsible for Statistics Canada



November 2007

Catalogue no. 89-552-XIE, no. 18

ISSN 1480-9516

ISBN 978-0-662-47311-4

Frequency: Irregular

Ottawa

Cette publication est disponible en français (N° 89-552-MIF au catalogue, n° 18)

Statistics Canada

# Acronyms

| | |
|---|---|
| BA | Bachelor of arts |
| IALS | International Adult Literacy Survey |
| IALSS | International Adult Literacy and Skills Survey |
| LFS | Labour Force Survey |
| NLSCY | National Longitudinal Survey of Children and Youth |
| OLS | Ordinary Least Squares |

## Acknowledgements

This report was made possible by the financial support of Statistics Canada. We are grateful to Scott Murray for encouraging us to undertake this project, and to a number of reviewers for their comments: Yvan Clermont and Scott Murray of Statistics Canada; Urvashi Dhawan-Biswal and Chris Li of the National Policy Research Team at Human Resources and Social Development Canada; and reviewers at Citizenship and Immigration Canada. We also thank Danielle Baum of Statistics Canada for her work in preparing the report.

## Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

# Table of contents


| | | |
|---|---|---|
| Acronyms | | 4 |
| Acknowledgements | | 5 |
| 1. | Introduction | 8 |
| 2. | Data | 10 |
| 3. | The distribution of literacy | 12 |
| 4. | The generation of literacy skills | 15 |
| | 4.1 Schooling and background effects | 15 |
| | 4.2 Age and generational effects | 19 |
| | 4.3 Post-school, work related literacy acquisition | 32 |
| 5. | The impact of literacy on earnings | 35 |
| | 5.1 A simple theoretical framework | 35 |
| | 5.2 Estimation results | 37 |
| 6. | Conclusion | 41 |
| References | | 42 |
| Appendix A | Education definitions in the IALS94, IALSS03, and LFS | 43 |
| Appendix B | First stage results for years of schooling | 45 |
| Endnotes | | 46 |

# Table of contents

## Charts


| | | |
|---|---|---|
| Chart 1 | Document literacy scores, 2003 | 12 |
| Chart 2 | Document literacy, age 26 to 35 | 20 |
| Chart 3 | Document literacy, age 36 to 45 | 21 |
| Chart 4 | Document literacy, age 46 to 55 | 21 |
| Chart 5 | Document literacy, age 56 to 65 | 22 |
| Chart 6 | Document literacy, less than high school | 22 |
| Chart 7 | Document literacy, high school | 23 |
| Chart 8 | Document literacy, post-secondary non-university | 23 |
| Chart 9 | Document literacy, university | 24 |
| Chart 10 | Document literacy, age 26 to 35 in 1994 | 25 |
| Chart 11 | Document literacy, age 36 to 45 in 1994 | 25 |
| Chart 12 | Document literacy, age 46 to 55 in 1994 | 26 |
| Chart 13 | Document literacy, age 56 to 65 in 1994 | 26 |
| Chart 14 | Document literacy, less than high school, age 26 to 45 in 1994 | 27 |
| Chart 15 | Document literacy, high school, age 26 to 45 in 1994 | 28 |
| Chart 16 | Document literacy, university, age 26 to 45 in 1994 | 28 |


## Tables


| | | |
|---|---|---|
| Table 1 | The joint distribution of literacy and income | 13 |
| Table 2 | Characteristics associated with low and high literacy | 14 |
| Table 3 | Log of document literacy regressions | 17 |
| Table 4 | Pooled regressions including cohort effects | 29 |
| Table 5 | Pooled quantile regressions with cohort effects, by education group | 31 |
| Table 6 | Regressions including literacy use at work and occupation variables | 33 |
| Table 7 | Earnings regressions | 38 |
| Table 8 | Quantile earnings regressions | 39 |
| Table A.1 | Education category comparisons: IALS and IALSS | 44 |
| Table A.2 | Education category comparisons: LFS | 44 |
| Table B.1 | Log literacy regressions | 45 |

# 1. Introduction

Adult literacy skills are of both fundamental and instrumental importance. Sen (1999) argues that we should aim for a society in which every person has the capability to pursue any reasonable version of what they perceive to be good. To do so requires both access to at least minimal levels of resources and possession of characteristics that Sen calls "functionings". One of the most important of these functionings is literacy. Without literacy, individuals cannot take a full and equal role in social and political discourse: they become less than equal members of society without the basic tools required to pursue their goals (Sen (1999)). Thus, in any attempt to build a better society, the distribution and generation of literacy is of fundamental importance. Literacy is also potentially important for instrumental reasons. An individual who improved his or her literacy might plausibly be expected to have better employment opportunities and command higher earnings, leading to a higher level of well being. From a societal point of view, a more literate workforce may be better positioned to adjust to change and to adopt new technologies. Thus, improving literacy for individuals may have spill-over effects on the productivity of the economy as a whole.

In this paper, we first examine the distribution of literacy skills in the Canadian economy and how they are generated. In large part, the generation of those skills must have to do with formal schooling and parental inputs into their children's education. We examine those issues, though not as completely as would be possible with a longitudinal dataset that includes literacy type questions, such as the NLSCY. We also investigate the nature of literacy generation in the years after individuals have left formal schooling and are in the labour market. Once we have established the core facts about literacy in the economy, we turn to examining the impact of increased literacy on individual earnings. We investigate both the causal impact of literacy on earnings and the joint distribution of literacy and income, arguing that the latter provides a more complete measure of Sen's notion of how well an individual is able to function in society. Thus, we discuss literacy's role in Canadian society both in the fundamental and the instrumental sense.

The key to our investigations, of course, is the data. We make use of two versions of a remarkable dataset that combines answers to demographic and labour market questions with literacy test scores. The main data we focus on is from the Canadian component of the 2003 International Adult Literacy and Skills Survey (IALSS). This is a very large survey with over 22,000 respondents, including an over-sample of people with a First Nations background. We examine the latter group in detail in another paper, excluding both them and immigrants from the analysis here. We also make use of the Canadian component of the 1994 International Adult Literacy Survey (IALS) in order to obtain a more complete picture of how literacy changes with age and across birth cohorts. To avoid confusion, we will refer to the 2003 dataset as IALSS 2003 and the 1994 dataset as IALS 1994.

Our investigations of literacy using IALSS 2003 yield several strong results. Literacy increases strongly (though at a decreasing rate) with years of schooling. Parental education levels also have a strong positive impact on literacy, with mother's education being particularly important. On the other hand, parental occupations do not have either an economically substantial or a statistically significant impact on literacy once we control for parents' education. Moreover, whether a respondent's mother worked when the respondent was 16 has no impact on the respondent's literacy level. Perhaps most interestingly, we find little relation between age or labour market experience and literacy. We found the same, strong result in earlier work with IALS 1994 (Green and Riddell (2003)). At first glance, this result appears to imply that individuals acquire their literacy through formal schooling and through the efforts of their parents but that their literacy levels are essentially "locked in" upon leaving school. However, using a combination of IALS 1994 and IALSS 2003, we show that the flatness of literacy relative to age in the cross-sectional datasets actually arises from a combination of offsetting ageing and cohort effects. In particular, individuals from a given birth cohort actually lose literacy skills in the years after they leave school. At the same time, we find strong evidence that more recent birth cohorts have lower levels of literacy. This is particularly true for more highly educated individuals and shows up mainly in a thinner right tail of the literacy distribution (i.e., in fewer people attaining high literacy scores). Thus, a 35 year old in IALSS 2003 has approximately the same average literacy score as a 25 year old in the same survey not because that 25 year old should expect to be at the same literacy level in 10 years but because the 35 year old started from a higher literacy level at age 25 (i.e., comes from a more literate cohort) but lost some of their initial literacy skills during the time since they left school. These results suggest, on the one hand, a tendency for literacy skills to decline over time and on the other that we are doing a poorer and poorer job of educating successive generations. By taking into account use of reading, writing and mathematical skills in the workplace, we also investigate whether literacy skills display a "use it or lose it" feature.

# 2. Data

Our data comes from the International Adult Literacy and Skills Survey (IALSS 2003): a fascinating survey carried out in several countries in 2003.[1] We also make use of the International Adult Literacy Survey (IALS 1994), an earlier survey of literacy skills carried out, for Canada, in 1994. The IALSS 2003 includes standard questions on demographics, labour force status and earnings, but it also attempts to measure literacy and related cognitive skills in four broad areas: Prose, Document, Numeracy, and Problem Solving (the latter is not included in the earlier IALS 1994). Perhaps of most importance for our discussion, the IALSS 2003 and the IALS 1994 did not attempt to just measure abilities in math and reading but tried to assess capabilities in applying skills to situations found in everyday life. Thus, the Prose questions in the surveys assess skills ranging from items such as identifying recommended dosages of aspirin from the instructions on an aspirin bottle to using an announcement from a personnel department to answer a question that uses different phrasing from that used in the text. The Document questions, which are intended to assess capabilities to locate and use information in various forms, range from identifying percentages in categories in a pictorial graph to assessing an average price by combining several pieces of information. The Numeracy component ranges from simple addition of pieces of information on an order form to calculating the percentage of calories coming from fat in a Big Mac based on a table. It is important for the work that follows that the Numeracy component changed substantially between the 1994 and 2003 surveys in response, in part, to concerns that the way the questions were stated in 1994 meant that math skills could not be separated from reading skills. In contrast, the Document and Prose tests have substantial overlap in the two survey years, with approximately 45% of the questions being identical across years. Statistics Canada also renormalized test results from the remaining 55% of the questions in 2003 so that the overall average test scores from 2003 bore the same relationship to the overall average in 1994 as do the averages on the questions that are identical between the two years. Based on this, in the analysis that follows, we treat the Prose and Document test scores as perfectly comparable between the two survey years. We subsequently discuss this assumption and evidence in favour of it.

The IALS 1994 sample contains observations on 5,660 individuals while IALSS 2003 is substantially larger at 23,038 individuals. Our goal is to focus on literacy generation in the Canadian economy and, as a result, we exclude from our sample anyone born outside of Canada from both samples in order to focus attention on the Canadian educational system. We also drop the over-sampled First Nations observations from the 2003 survey, reserving a more careful examination of those individuals for another paper. The surveys cover individuals over age 16 but we exclude individuals who list their main activity as student in order to highlight the effect of completed schooling and what happens to literacy afterwards. The result is samples of size 3,964 for the IALS 1994 and 14,666 for the IALSS 2003, which form the basis of our initial analysis of the distribution of literacy skills in Canadian society. However, when we turn to our investigation of the generation of literacy after leaving school and to the impact of literacy on earnings, we restrict ourselves to a sample of individuals in which we eliminate the self-employed and workers with weekly earnings that are less than \$50 and over \$20,000. The latter restriction eliminates retired people, the unemployed and others who are not in the labour force. It also cuts out a small number of individuals with earnings that are substantial outliers relative to the rest of the sample. We drop the self employed because we wish to examine the remuneration of skills in the labour market, and self employed earnings reflect both that remuneration and returns to capital. We include both males and females throughout, dividing the analysis on gender lines in some places. Finally, we use the sample weights provided with the data in all tables and estimation.

In the earnings analysis, we use a constructed weekly earnings variable. In the IALSS 2003, respondents are first asked about their standard pay period and then asked about typical earnings in that pay period. From those answers, we construct a weekly earnings variable for each person. Thus, for individuals who report that they are paid monthly, we divide their reported earnings by 4.333.

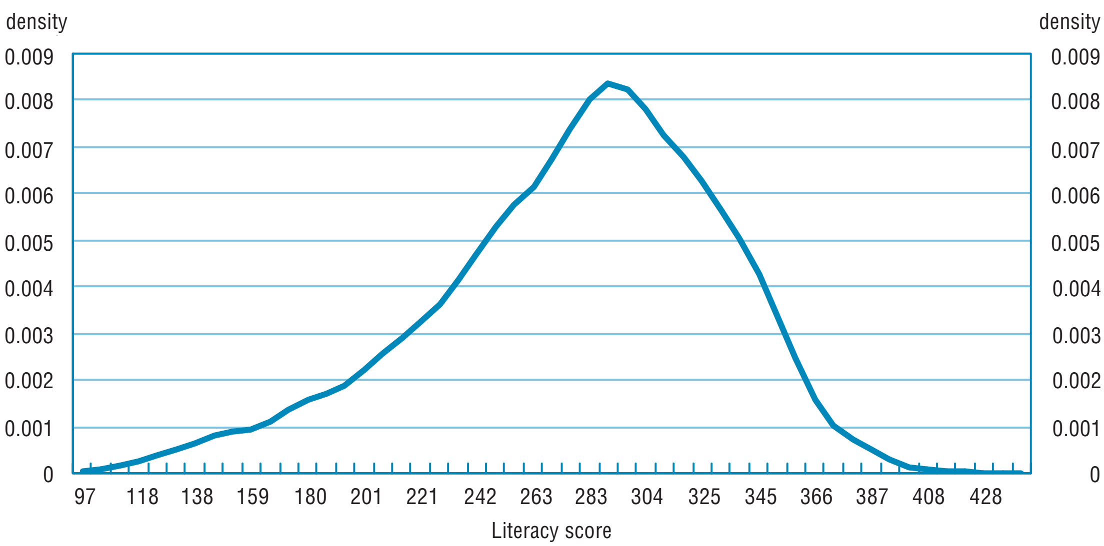
One defining feature of this data is the strong correlation among the various literacy scores. Thus, the correlation between the Prose and Document scores in the IALS 2003 is 0.96, the correlation between Document and Numeracy is 0.92, and the correlation between the Document and Problem Solving scores is 0.92. Further, a principal components analysis indicated essentially one principal component placing equal weight on all four scores. Thus, a simple average of the scores captures much of the information available in the four scores. Because the scores are so highly correlated, we will work with one specific score in our description of the literacy distribution and generation. We prefer to focus on a single meaningful variable rather than an amalgam of scores in this part of the analysis because it makes cleaner the discussion of how the score is generated. More specifically, we focus on the Document Score. We need to choose either the Document or Prose scores because they are the ones that are common across IALS 1994 and IALSS 2003, and the Document score also happens to be the one with the highest associated factor loading in the factor analysis. None of the conclusions we present below would change if we presented the discussion in terms of one of the other scores. When we turn to the earnings analysis, where we investigate the overall effect of literacy we use a simple average of the four scores. This also allows a clearer comparison with our earlier work.

# 3. The distribution of literacy

We begin with a simple description of the distribution of Document literacy based on the IALSS 2003. In Chart 1, we present a kernel density plot for Document literacy for our sample. The density is slightly negatively skewed with a mean of 280, a median of 287 and a standard deviation of 50. The fact that the distribution is negatively skewed when ability is typically thought of as distributed according to a positively skewed distribution such as the log normal may be a reflection of a lack of sufficient questions to differentiate among highly literate respondents. The minimum Document score in the sample is 84, representing extremely low proficiency at the Document test tasks, while the highest is 436.[2] Inequality in the distribution is reflected in an associated Gini coefficient value of 0.107 and a value for the log of the ratio of the 90th percentile to the 10th percentile of 0.513. To put this in perspective, the Gini for pre-tax and transfer family income from the 2001 Canadian Census is 0.438 and the log of the 90 : 10 ratio is 3.48 (Frenette et al, 2006). Thus, the distribution of literacy is much less unequal than that of market income. This is not surprising since literacy skills are only one of the components playing a role in earnings generation.

**Chart 1**

**Document literacy scores, 2003**



The other comparison we can make with this data is with the extent of inequality in the 1994 literacy distribution. The Gini for that distribution is 0.151 and the log of the 90 : 10 ratio is 0.76. Thus, the Document literacy distribution became substantially less unequal in the nine years between the surveys. In part, this occurred because of an improvement at the low end of the literacy distribution. The 10th percentile of the 1994 literacy distribution was 160, compared to a value of 197 in 2003. The news is not unambiguously good, however, since the 95th percentile of the distribution actually fell from a value of 359 in 1994 to 351 in 2003. Thus, the decline in literacy inequality reflects both substantial improvements at the low end of the distribution and some deterioration at the top end. We will return to discussing these differences below, but for the moment we will focus on characterizing the 2003 distribution.

In Table 1, we present a depiction of the joint distribution of literacy and family income, showing the percentage of observations in each cell in a grid defined by the 5 quintiles of the Document literacy distribution and the five quintiles of the household (pre-tax and transfer) income distribution. The correlation between household income and literacy equals 0.31, indicating that the two variables are positively correlated, but not strongly so. This is seen in Table 1 from the fact that cells on the main diagonal are not overly large relative to the other entries. Perhaps the most interesting cell is the one corresponding to individuals in the bottom quintile of both the literacy and income distributions – a cell which contains just over 8% of the population. These individuals could be viewed as suffering from a double poverty – both in terms of literacy and income. Everyone in this cell is what Crompton (1996) describes as “marginally literate”. In terms of document literacy, Crompton describes the marginally literate as unlikely to be able to complete tasks such as using a bus schedule to chart out what time the last bus leaves a particular stop on Saturday night (Crompton (1996)). Thus, they are severely disadvantaged in terms of their ability to function in society. Perhaps not surprisingly, these doubly poor individuals have extremely low education levels, with 80% of them being high school drop outs. The same pattern can be seen in their family background, with over 90% of them having fathers who were also drop outs. In contrast, the 7.4% of individuals who are in both the top quintile for literacy and the top quintile for income are very highly educated and come from high education backgrounds. Fully 30% of this group are university graduates and an additional 21% have post-graduate education. One-quarter of their fathers also completed university, which is much higher than for the population as a whole, in which just over 9% of fathers had a university degree.

**Table 1**
**The joint distribution of literacy and income**

| | Income quintile | | | | |
|---|---|---|---|---|---|
| Literacy quintile | 1st | 2nd | 3rd | 4th | 5th |
| | | | percent | | |
| 1st | 8.3 | 5.8 | 3.4 | 1.6 | 0.8 |
| 2nd | 4.4 | 5.3 | 4.4 | 3.4 | 2.4 |
| 3rd | 2.5 | 3.9 | 4.8 | 4.9 | 4 |
| 4th | 1.9 | 3.2 | 4.4 | 5.2 | 5.4 |
| 5th | 1.3 | 2.4 | 4.2 | 4.9 | 7.4 |

**Note:** Based on calculations from IALS 2003. Literacy refers to Document Literacy. Income refers to household pre-tax and transfer income. Cell entries correspond to percentage of total observations in the given cell, defined by quintiles of the literacy and income distributions. Numbers in each row and column may not add up to 20 due to rounding.

To provide a further perspective on potential difficulties associated with low literacy, in Table 2 we present measures of societal participation and individual well-being for low and high literacy individuals. We define a low literacy group as a set of people who were recorded as level 1 when we use all five plausible Document test scores. In terms of literacy skills, this is an extremely disadvantaged group of individuals who are unlikely to be able to do tasks such as process patterns in more than one chart (Crompton ,1996). Our high literacy group consists of individuals scoring in either of the top two literacy levels. The low literacy group claims to have been more likely to have voted in a municipal election than the high literacy group, though this is the only measure of participation in which they lead. Indeed, the high literacy group was more likely to have voted in a federal or provincial election, though the difference is not large. Differences are large, however, in participation in political organizations and in community and school groups, with the high literacy group being at least twice as likely to participate in either of these groups. We are not claiming that the link between literacy and participation is causal (i.e., that increasing a group's literacy would result in an increase in their participation) but the strong correlation does emphasize a point made by Sen (1999) that literacy bears an important relationship to full participation in society. This is evident as well in the fact that nearly 11% of low literacy individuals stated that having to make basic calculations such as calculating a tip makes them anxious. The strongest association, though, is in terms of self-assessed health. Approximately 76% of high literacy individuals rated themselves as being in either excellent or very good health compared to only 30% among the low literacy group. To make sure this is not just a reflection of poor health causing individuals to have trouble completing the tests, these percentages are calculated after dropping individuals who were listed as not completing the basic questionnaire for reasons related to health or disability. In part, the health-literacy differential reflects the fact that the low literacy group is much older, revealing generational differences in literacy to which we return below. However, even if we drop individuals over age 65 and estimate a linear probability regression of a dummy variable equalling one if an individual rated herself as in very good or excellent health on age, years of schooling, family income and a dummy variable corresponding to people in the low literacy category, the latter variable has a coefficient of -0.16 and is highly statistically significant (with an associated standard error of 0.041). Thus, low literacy individuals are approximately 16% less likely to claim they are in good health than everyone else, even after controlling for age, years of schooling and family income. Again, whether this is causal is not clear but it does raise serious concerns about the functioning of low literacy individuals in Canadian society.

**Table 2**
**Characteristics associated with low and high literacy**

| Variable | Low literacy group | High literacy group |
|---|---|---|
| | percent | |
| Voted municipal | 67.9 | 61.7 |
| Voted federal | 79.5 | 84.2 |
| Participation – political organization | 2.6 | 6.7 |
| Participation – community / school | 11.4 | 47.4 |
| Satisfied with life | 78.3 | 82.5 |
| Good health | 29.9 | 75.7 |
| Anxious about numbers | 10.7 | 3.8 |
| Household income (Median) | 30,000 | 90,000 |

**Notes:** Calculations from IALS 2003. All entries are percentages except for household income which is in dollars. Low literacy group are individuals who scored in level 1 in all 5 Document Literacy tests. High literacy group are individuals who scored in levels 4 or 5 in all 5 Document Literacy tests.

# 4. The generation of literacy skills

## 4.1 Schooling and background effects

We turn now to using both IALS 1994 and IALSS 2003 to examine the sources of literacy. We do this using a series of regressions, where we again focus on Document literacy as our representative measure. More specifically, we use the log of the Document literacy score as our dependent variable so our estimated coefficients can be interpreted as showing impacts in terms of percentage changes in literacy. We begin with the much larger IALSS 2003, examining the impacts of family background and schooling on individuals' literacy scores.

Before presenting the estimation results, we begin by setting out a brief, heuristic model of literacy generation. The model will help to put our estimates in context as well as providing guidance in thinking about identification issues. Consider a simple model in which individuals start out at birth endowed with two key characteristics: their ability and parental resources. By parental resources, we mean something quite broad, incorporating both parental income and parental willingness and ability to support their children's education and literacy acquisition. Pre-school children begin to acquire literacy based on these fundamental characteristics (ability and parental resources). Once they enter school, these characteristics interact with characteristics of the school such as teacher quality, class size and the attitudes and abilities of peers. New additions to literacy with each year of schooling are then functions of ability, parental resources, school characteristics and the literacy level at the beginning of the period. These influences may interact in complicated ways. These additions continue until the legal school leaving age. After that point until the end of high school, students make a decision each year on whether to continue in school. That decision will be a function of ability, parental resources and school characteristics, again, but it is also likely to be a function of literacy acquired to that point. The more literate a student is, the less onerous they are likely to find school and, thus, the more likely they are to choose to stay an extra year. Finally, after high school, whether an individual continues to go to school will be determined by a combination of their own decision to apply to continue and the decision of the college or university on whether to admit them. The latter decision will likely be a function of the student's literacy as reflected in her grades. Thus, schooling and literacy are co-determined with extra years of schooling leading to increased literacy but increased literacy also leading to more years of schooling, especially after the legal school leaving age. Indeed, once we account for expectations, the inter-relation between the two may be even tighter. Individuals who do not expect to continue with school past the legal minimum may rationally under-invest in acquiring literacy skills while they are in school.

Once individuals leave school, literacy acquisition is likely more difficult. Literacy skills may be acquired on the job if they are needed for carrying out tasks at work but otherwise further acquisition would require active investment in non-work hours. Indeed, it seems quite possible that individuals could lose literacy skills after they leave formal schooling if those skills depreciate when they are not used.

We are interested in characterizing as many of the components of literacy generation as possible. In particular, we are interested in the relationship of literacy to parental resources since that relationship is fundamentally linked to notions of equity: to the extent that one generation's literacy hinges on the resources of the previous generation, differences in literacy can be seen as arising from characteristics beyond the control of the people involved. This is a paradigmatic case for redistributional policy in the views of philosophers such as Dworkin and Rawls. We are also interested in the relationship between formal schooling and literacy since this is a main channel through which we could hope to affect the literacy distribution. Finally, we are interested in whether literacy declines or improves after leaving school and how this process is related to characteristics of an individual's job. If literacy has a "use it or lose it" form then there may be good reason to adopt policies such as subsidizing firms to provide access to literacy maintenance activities such as allowing employees to return to school. Many of these parameters of interest reflect causal relationships that are difficult to establish definitively. We will make efforts to estimate the causal parameters where the data permit but much of what we will discuss is necessarily in the form of correlations rather than clear causal impacts.

In the OLS 1 column of Table 3, we present our simplest OLS regression: one in which the dependent variable is the log of individual Document literacy and the independent variables are age, age squared, years of schooling, years of schooling squared, and a gender dummy. Thanks to the substantial sample size, all the variables are statistically significantly different from zero at either the 1% or 5% level but this does not mean their actual impacts are sizeable. Thus, the estimates show that women have lower literacy than men (conditional on school and age) but only by 1.4%. Similarly, the age and age squared coefficients are highly statistically significant but together they imply that the impact of an extra year of age on literacy is actually -0.1% at age 30.[3] This finding that there is essentially no relationship between literacy and either age or experience is a key part of the discussion in Green and Riddell (2003). The one relationship that is economically substantial is the one between literacy and schooling. One extra year of schooling, evaluated when the individual already has 12 years of education, increases literacy by 3.2%. This is very similar to what Green and Riddell (2003) calculated using the IALS 1994.[4]

In the OLS 2 column of Table 3 we add variables on parental education and immigrant status to our OLS regression. Introducing these variables has virtually no impact on either the gender or schooling variable effects. However, including them leads to a near doubling in the age coefficient. Given that the coefficient on the age squared variable also becomes more negative, the net effect of age is still quite small. The parental education variables are jointly highly significantly different from zero but, perhaps surprisingly, the effect is found almost entirely at low levels of parental education. Having a parent (either mother or father) who is a high school drop out, decreases average literacy by between 3% and 4%. However, parental education beyond high school graduation has no further impact on literacy. Interestingly, not knowing a parent's education level (which is the case for approximately 8% of the sample) has a strong effect, being associated with approximately 6% lower literacy. While we just included this variable in order to allow us to keep the observations for which parental education is missing, it seems possible it is actually proxying for something real. For example, children who do not know a parent's education likely did not have a close relationship with that parent. Thus, the estimated coefficient may reflect the extent to which literacy is generated through direct parental involvement. Finally, having a father who is an immigrant has a mild association with literacy (increasing literacy by 1.5%), while having a mother who is an immigrant has no impact. We also tested specifications in which we included a set of parental occupation dummy variables, but these were never jointly statistically significant. In particular, a test of the hypothesis that the set of father's occupation dummy variables jointly had zero effects has an associated P-value of 0.13. The same test for mother's occupation has a P-value of 0.79. We also find that a dummy variable representing whether the individual's mother was working when the individual was 16 does not have a statistically significant effect. Overall, the results point to a surprisingly weak association between literacy and parental background. Only schooling seems to have a substantial impact on literacy generation.

**Table 3**

**Log of document literacy regressions**

| Variable | OLS 1 | Standard error | OLS 2 | Standard error | OLS 3 | Standard error | OLS 4 | Standard error | IV | Standard error |
|---|---|---|---|---|---|---|---|---|---|---|
| Female | -0.014** | (0.0046) | -0.012** | (0.0045) | -0.0083 * | (0.0045) | -0.012** | (0.0044) | -0.016** | (0.0051) |
| Years of schooling | 0.058*** | (0.0053) | 0.054*** | (0.0049) | 0.053*** | (0.0048) | 0.024*** | (0.0008) | 0.052*** | (0.013) |
| Schooling squared | -0.0011*** | (0.0002) | -0.0011*** | (0.0002) | -0.0011*** | (0.0002) | ... | ... | ... | ... |
| Age | 0.0035*** | (0.0008) | 0.0064*** | (0.0008) | 0.0066*** | (0.0008) | -0.0068*** | (0.0008) | -0.0003 | (0.0036) |
| Age squared | -0.0076*** | (0.0009) | -0.0099*** | (0.0009) | -0.011*** | (0.0009) | -0.011*** | (0.0009) | -0.0029 | (0.0039) |
| **Mother's education** | | | | | | | | | | |
| Less than high school | ... | ... | -0.037*** | (0.0059) | -0.038*** | (0.0058) | -0.028*** | (0.0058) | 0.0007 | (0.016) |
| Some post secondary | ... | ... | -0.0077 | (0.0067) | -0.0084 | (0.0066) | -0.0046 | (0.0065) | -0.020 * | (0.011) |
| BA or more | ... | ... | 0.0094 | (0.0102) | 0.0075 | (0.0098) | 0.0076 | (0.01) | -0.0050 | (0.016) |
| None reported | ... | ... | -0.067*** | (0.0123) | -0.069*** | (0.0122) | -0.048*** | (0.011) | -0.0047 | (0.024) |
| **Father's education** | | | | | | | | | | |
| Less than high school | ... | ... | -0.032*** | (0.0067) | -0.03*** | (0.0067) | -0.029 | (0.0067) | -0.008 | (0.012) |
| Some post secondary | ... | ... | 0.0055 | (0.0073) | 0.006 | (0.0071) | 0.0062 | (0.0073) | -0.008 | (0.012) |
| BA or more | ... | ... | 0.0126 | (0.0083) | 0.0154 * | (0.0082) | 0.011 | (0.0082) | -0.025 | (0.02) |
| None reported | ... | ... | -0.056*** | (0.012) | -0.055*** | (0.012) | -0.065*** | (0.011) | -0.021 | (0.023) |
| Immigrant mother | ... | ... | 0.0056 | (0.0077) | 0.0053 | (0.0076) | -0.0009 | (0.0077) | -0.013 | (0.011) |
| Immigrant father | ... | ... | 0.015** | (0.0071) | 0.016** | (0.007) | 0.0054 | (0.0071) | -0.0084 | (0.011) |
| Good math grades | ... | ... | ... | ... | 0.027*** | (0.0055) | ... | ... | ... | ... |
| Teachers too fast | ... | ... | ... | ... | -0.026*** | (0.0057) | ... | ... | ... | ... |
| Constant | 5.08*** | (0.04) | 5.09*** | (0.038) | 5.09*** | (0.038) | 5.29*** | (0.02) | 5.04*** | (0.12) |
| Observations | 14,527 | ... | 14,527 | ... | 14,527 | ... | 13,868 | ... | 13,868 | ... |
| R-squared | 0.49 | ... | 0.51 | ... | 0.52 | ... | 0.47 | ... | 0.310 | ... |

... not applicable
* Statistically significant at 10% level.
** Statistically significant at 5% level.
*** Statistically significant at 1% level.
**Note:** The instruments for schooling in the last column are province of high school attendance dummy variables and the latter variables interacted with age.

As we discussed earlier, literacy and years of schooling are likely to be jointly determined. In that case, our OLS coefficient on schooling provides a biased estimate of the impact of schooling on literacy. We attempt to address this in two ways. First, biases may arise because of a correlation between literacy and schooling arising from unobserved ability. If high ability people do not view it as particularly costly to either acquire literacy or go to school then we could observe a strong, positive coefficient on schooling in our regression because years of schooling is proxying for ability rather than as a reflection of a causal impact of schooling on literacy. This problem can be addressed if we have a measure of ability since once we control for ability, any relationship between schooling and literacy cannot be due to an omitted ability term. Note, though, that many studies that try to control for ability (in, for example, earnings regressions) actually use scores on tests much like our literacy tests. What we would require is a test score from a very young age – before the process we are trying to study really begins. Since we don't have that, we instead try to proxy for ability using two variables that are plausibly related to it. In particular, in the third regression, we include a dummy variable equalling one if the person agreed or strongly agreed with the statement that they got good grades in math when they were in school and another dummy variable equalling one if the respondent agreed or strongly agreed with the statement that teachers often went too fast and the person often got lost. Either of these could plausibly be seen as proxies for innate ability. Both of these variables enter significantly, with people who claimed to have gotten good grades in math having 2.7% higher literacy and those who thought teachers went too fast having 2.6% lower literacy. However, including these variables has almost no impact on the other estimated coefficients. Most notably, their inclusion does not change the impact of schooling on literacy at all.

An alternative approach to the problem of identifying a causal effect is to find an instrumental variable for schooling and estimate our regression using two-stage least squares. An instrument is a variable which plausibly affects the right hand side endogenous variable directly but affects our dependent variable only through its relationship to the right hand side endogenous variable. Essentially, it allows us to get estimate effects based on the part of the variation in the right hand side endogenous variable that is correlated with the instrument and, hence, uncorrelated with the error term. In our case, this permits consistent estimates of the true impact of schooling on literacy providing two conditions are met: 1) the instrument truly does affect education; and 2) the instrument does not belong in the regression determining our dependent variable in its own right. We use the province where the individual resided when he or she was last in high school or primary school fully interacted with age as an instrument for schooling. The idea behind the instrument is that different levels of public resources applied to schooling in different provinces for different generations will lead to different schooling outcomes for otherwise identical individuals. On the other hand, we cannot see a reason for province of residence having an impact on literacy other than through changing schooling outcomes. In implementing this approach it is important that we also control for current province of residence in both the first stage (schooling regression) and the second stage (literacy regression). Province of current residence may be related to literacy if more literate individuals choose to migrate to provinces with a higher proportion of high literacy jobs and low literacy individuals chose to move to, for example, provinces with large numbers of resource jobs. In that case, to the extent that province of residence during high school and current province of residence are correlated, the province of residence would pick up this migration effect rather than the schooling effect we want it to capture. Controlling for province of current residence addresses this problem and means that we are really identifying the schooling effect from people who currently reside in the same province but were schooled in different provinces at different times.

The results from our two stage least squares estimation using province of residence during high school as an instrument is given in the IV column. We are unable to come up with a separate instrument for schooling squared. A standard approach to instrumenting for a squared term would be to use higher order terms in the instrument as instruments for the squared term. However, since our instrument is a dummy variable (or, more specifically, a set of dummy variables), the higher order terms equal the dummy variable itself and we are left without a separate instrument for the squared term. As a result, we estimate a specification that just includes the linear schooling variable. In the OLS 4 column, we present estimates from an OLS regression using the same specification for comparison.[5]

In the first stage (reduced form) regression, in which years of schooling is the dependent variable, the set of covariates include the full set of covariates from our main literacy regression: age, age squared, parental education, parental immigrant status and province of current residence. It also includes the set of dummies representing the province at time of high school. As expected, parental education is strongly positively related to years of schooling. Having parents who are immigrants also corresponds to having more years of schooling. The set of provinces at time of high school variables and their interactions with age are jointly highly statistically significant, indicating that the requirement that the instrument is a significant determinant of the endogenous variable is satisfied.[6] The first stage estimates are presented in Appendix B.

In the OLS 4 column results, the estimated schooling effect is highly significant and of the order of magnitude we reported as the percentage impact of an added year of schooling on literacy when the individual has 12 years of schooling based on the quadratic specification. The estimate of the schooling effect presented in column IV, however, is approximately twice as large, implying even stronger schooling effects than we have estimated with OLS specifications. Interestingly, once we instrument for schooling, the parental background variables all become smaller in magnitude and lose statistical significance. This may indicate that the parental background effects we estimated earlier were actually disguised place-of-schooling effects or, at least, that the parental background and place-of-schooling variables are highly collinear. As in earlier specifications, gender and age continue to have small impacts on literacy. We do not present the coefficients corresponding to the provincial dummy variables for the sake of brevity but they show that the Atlantic provinces and Ontario have essentially similar literacy levels, Quebec has significantly lower literacy levels and the Prairies and BC all have significantly higher levels. Our main conclusion is that, if the assumptions underlying our instrument are correct, these results indicate that education has a strong causal effect on literacy and that schooling is the dominant determinant of literacy.[7] To put the estimated effect in perspective, completing four extra years of schooling (e.g., moving from being a high school graduate to a university graduate) implies a 21% increase in literacy, based on the IV estimates. This would be enough to move the individual from the median to approximately the 90th percentile of the literacy distribution in 2003.

## 4.2 Age and generational effects

One of the most striking results from our initial regressions is the lack of significance of the age variable. Taken at face value, this suggests that individuals neither lose nor acquire new literacy skills after they finish schooling. Essentially, literacy is acquired at school and then simply maintained thereafter. This finding is reinforced if we replace age and age squared by experience and experience squared in the first regression specification from Table 3.[8] The coefficients on schooling, schooling squared and the female dummy are virtually identical to those presented in Table 3. The coefficient on experience is very small: 0.00051 with a standard error of 0.00045.

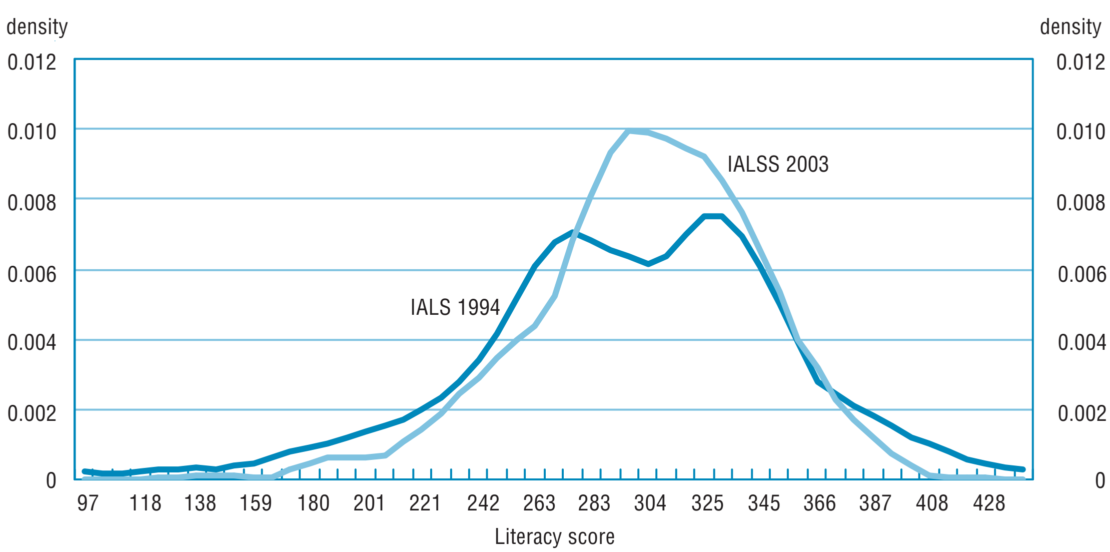
Interpreting the coefficient on age in a cross-sectional regression requires some care, however. As has been extensively studied in the literature on immigrant earnings, differentials between two age groups in a cross section could reflect a variety of possible combinations of true age and generational effects. Thus, while we are tempted to view the literacy level of 35 year olds in the IALS 2003 as a reflection of the literacy level the 25 year olds are likely to be at in 10 years time, we need to bear in mind that the 35 year olds come from an older generation and their observed literacy reflects a combination of any generational differential as well as any ageing effect. Only if there are no systematic differences across generations does the cross-sectional literacy–age profile reflect the true impact of ageing on literacy.

A more complete investigation of generational effects (often called cohort effects) and ageing effects requires the use either of true panel data or of at least two cross-sectional datasets constructed in such a way that we can follow "synthetic" cohorts through time. We make use of the IALS 1994 and the IALSS 2003 for this purpose. More specifically, in the public use version of the IALS 1994, we can observe a set of 10 year wide age groups for the respondents (i.e., ages 26 to 35, 36 to 45, 46 to 55, 56 to 65, and 65and over). Since we have a continuous age variable in the IALS 2003, we can construct age groups that correspond to the age people in these initial groups would be 9 years after IALS 1994 (i.e., 35 to 44, 45 to 54, 55 to 64, 65 to 74, and 75 and over). We call the set of people who were age 26 to 35 in 1994 and 35 to 44 in 2003, cohort 1 and number the remaining cohorts in ascending order from there. If both surveys provide random samples then each provides an unbiased estimate of the literacy distribution for the cohort at two different points in time and we can follow the progress of a given cohort over time. In part of our examination, we will further break the cohorts down by education. Linking across surveys would not provide unbiased estimates of the progress of cohort–education groups in this case if the educational composition of the groups being studied changes over time. Thus, we cannot say that the average literacy of 35 to 44 year old high school graduates in IALSS 2003 provides a consistent estimate of what happened to the 26 to 35 year old high school graduates we observe in IALS 1994 over the intervening 9 years if some of the initial high school graduates would be expected to have increased their education level. In that case, the average literacy in 2003 would reflect a combination of ageing effects and the systematic selection of people out of the high school graduates group. To avoid this problem, we focus on individuals over age 25, after which point changes in formal education across our broad categories are rare. It is also worth noting that the oldest cohort (made up of those over age 65 in 1994) is special in that its composition is likely to change over time because of deaths. As a result, while we include it in our analysis, we do not place much weight on results related to that cohort in our conclusions.

We begin with plots of the Document literacy densities from the IALS 1994 and IALSS 2003 broken down by age, education and cohort groups. Thus, in Chart 2, we present the density plots for individuals who are age 26 to 35 in 1994 and the individuals who are the same age in 2003. These densities correspond to the literacy for the youngest cohort we examine in 2003 and for the cohort just before them at the same age (observed in 1994). The density for the younger cohort (observed in 2003) has noticeably less spread, with both the left and right tails being thinner. In other words, the younger cohort experiences an improvement in literacy at the low end of the distribution but a worsening at the top end. This pattern is reflected in a 10th percentile of 223 and a 90th percentile of 363 in 1994 compared to values of 248 and 354, respectively, in 2003. The reductions in inequality in both tails nearly cancel out, as reflected in a median value of 299 in 1994 as compared to 306 in 2003. Thus, one might conclude from a measure of central tendency that little has changed, whereas the tails reflect more substantial differences.

**Chart 2**

**Document literacy, age 26 to 35**



In Chart 3, we present the same kind of density comparison but for individuals aged 36 to 45. The outcome is much the same as in Chart 2 except that the relative improvement in the lower end of the distribution is not as great and the relative decrease at the top is decidedly larger than what we observed in Chart 2. In contrast, Chart 4 shows that for the 46 to 55 year olds, the results in 2003 are better across the distribution. This is important, in part, because observing different relative changes between 1994 and 2003 for different groups suggests that we are witnessing something real rather than just a difference in the tests in the two years. If all plots for all groups showed improvements at the bottom and declines at the top between 1994 and 2003 then the simplest explanation for the patterns would be that the test changed in such a way that it generated better results on the easy questions that will constitute most of the scores at the bottom but worse results on the harder tests that will define the shape of the top of the distribution. Finally, Chart 5 also shows improvements in 2003 for the 56 to 65 age group, though the 1994 and 2003 densities are very similar at the very top end.

**Chart 3**

**Document literacy, age 36 to 45**

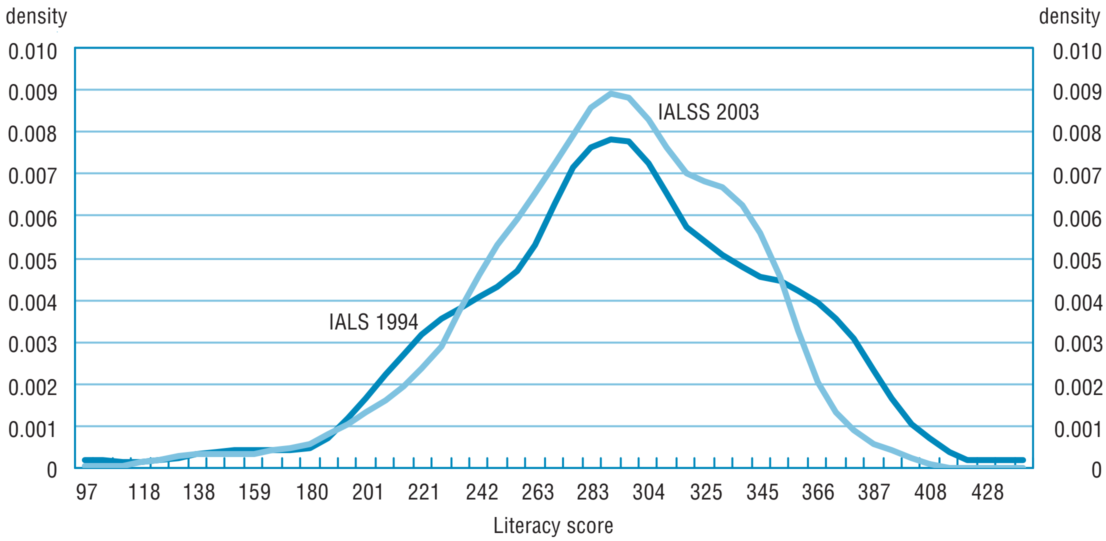



**Chart 4**

**Document literacy, age 46 to 55**

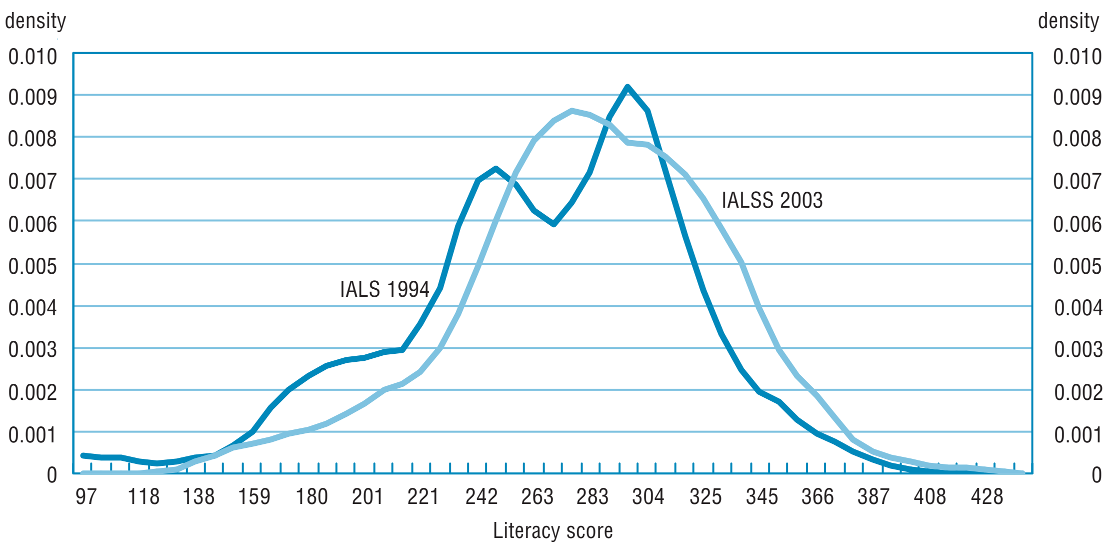
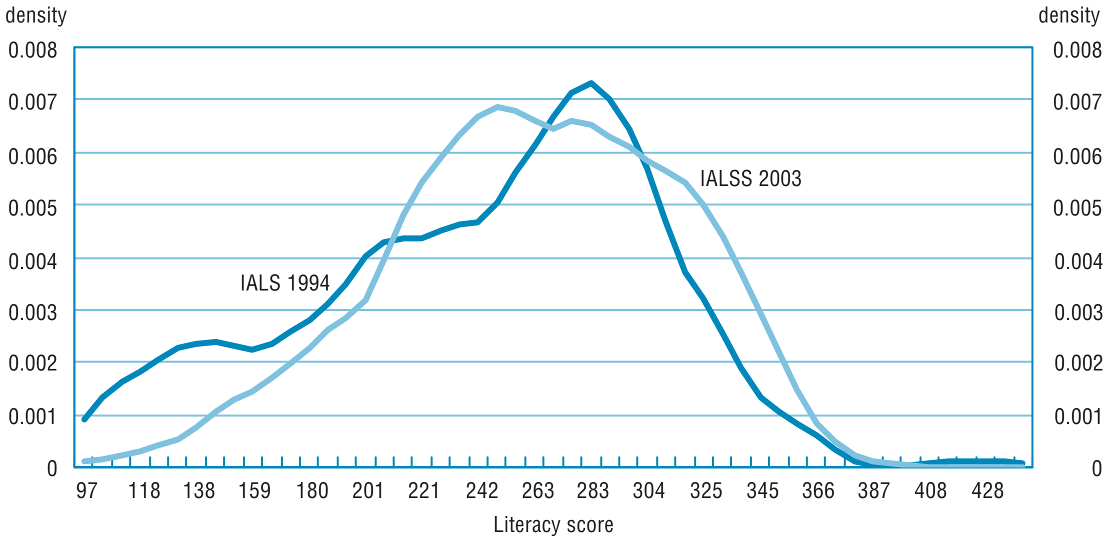


**Chart 5**

**Document literacy, age 56 to 65**




We next examine the changes between the two years broken down by education group. Thus, in Chart 6, we plot the Document literacy densities in 1994 and 2003 for people with less than a high school diploma. The improvement at the bottom end of the overall distribution is also evident at the bottom of this distribution and is quite large. The 10th percentile for this drop-out group rises from 145 in 1994 to 175 in 2003. At the other end, the differences are smaller. While the 95th percentile declines between 1994 and 2003, the 90th percentiles are actually identical in the two years. Of course, the 90th percentile of their distribution is relatively low – being around the 70th percentile of the overall distribution.

Even by the time we turn, in Chart 7, to high school graduates, however, there is clearer evidence of a worsening in the right tail of the distribution with a much smaller improvement in the left tail than was evident with the high school drop-outs.

**Chart 6**

**Document literacy, less than high school**

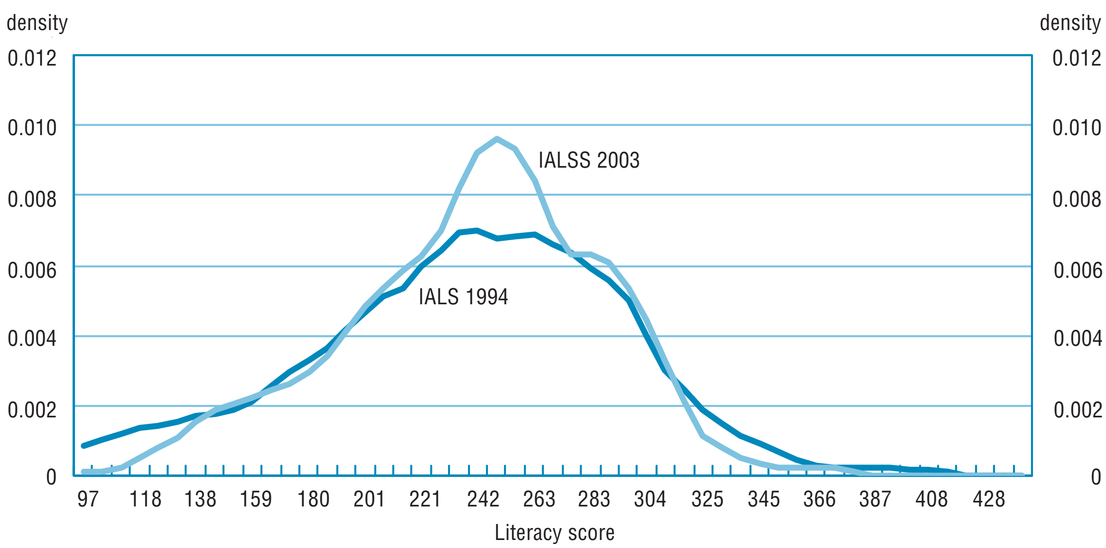
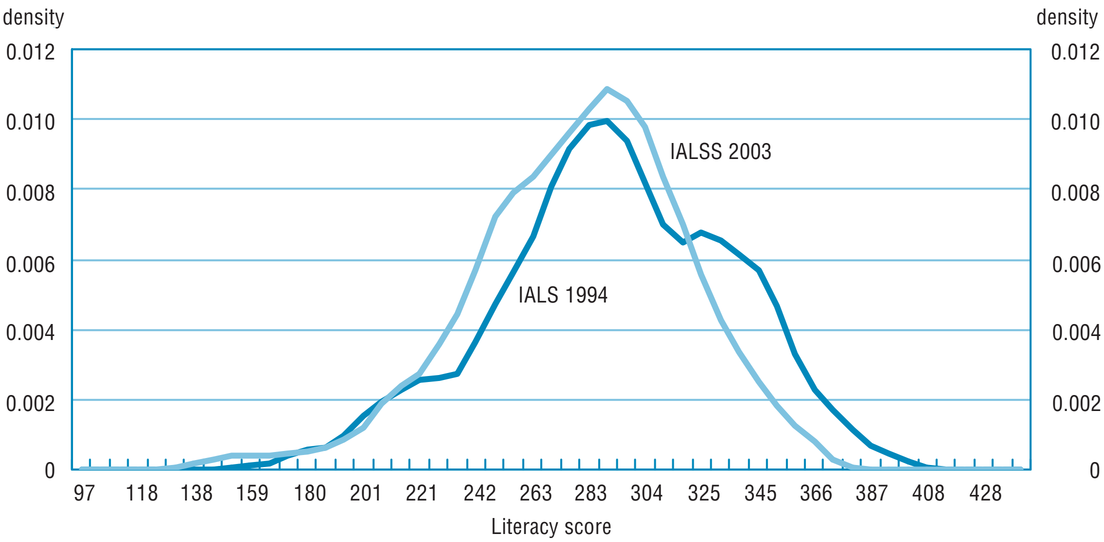


**Chart 7**
**Document literacy, high school**




Both the Non-university Post-Secondary and the University groups show, in Charts 8 and 9 respectively, declines in the upper part of their distributions, with the former group also showing some declines lower in the distribution as well. Both the Some Post-Secondary and the University groups register approximately 10% declines in their 90th percentiles between the two years (from 373 to 338 for the Post Secondary and from 392 to 359 for the University group). Thus, in terms of within education group movements, the improvements at the bottom of the overall distribution are really only clearly evident in the high school drop-outs distribution while the declines at the top are mainly observed in the distributions for the top two groups.

**Chart 8**
**Document literacy, post-secondary non-university**

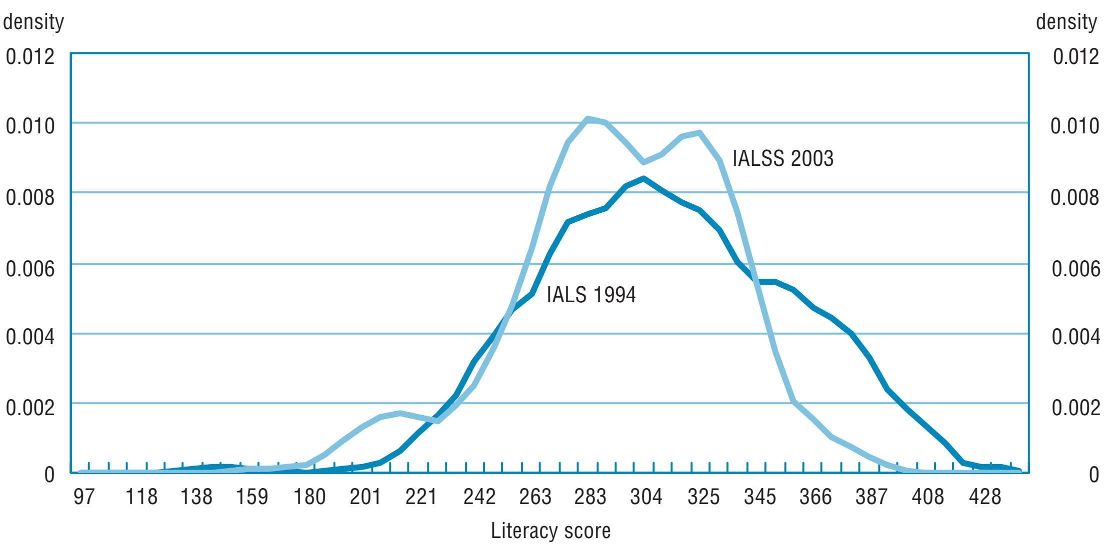
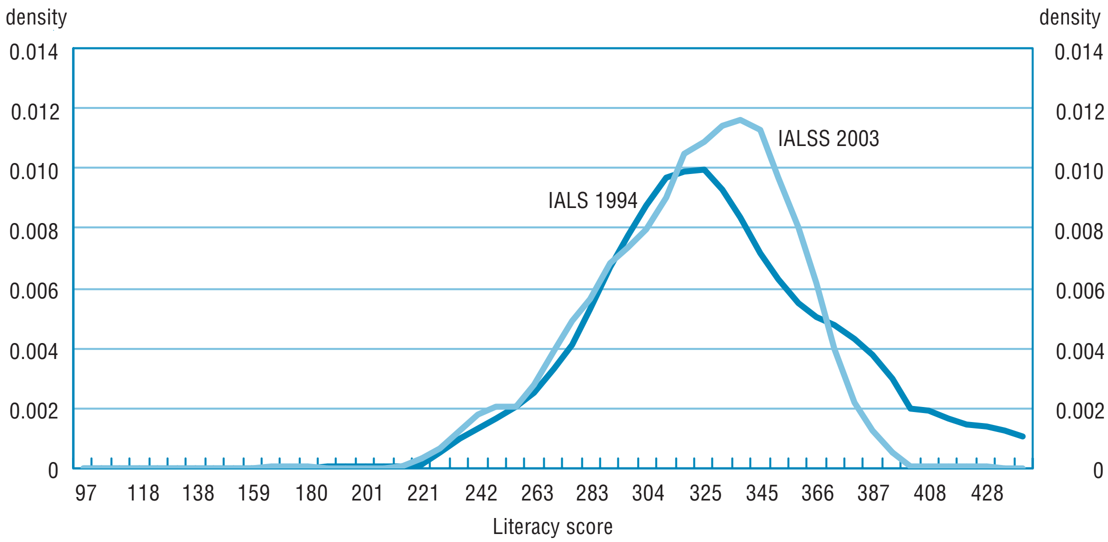


**Chart 9**

**Document literacy, university**




The changes observed for the overall distributions may, in part, also stem from changes in the educational composition of the population between the two years. The replacement of older, less educated cohorts with younger, more educated ones shifts the educational distribution. Thus, the proportion of the population with a Bachelor of Arts rises from 17% in 1994 to 21% in 2003. This improvement undoubtedly plays a role in the improvement in the lower tail of the overall distribution. It is difficult to see, however, how this change could cause the declines at the top of the distribution. If a person of the same underlying ability was a high school graduate in the oldest cohort but a university graduate in the newest cohort then if we were to replace the former with the latter, one would expect the overall distribution to either improve (if university education increases literacy) or stay the same (if it does nothing). Even the change at the top of the distribution for university graduates is unlikely to be related to an expansion of the university educated pool. While the additional university graduates (those who would not have attended university in an earlier generation) are likely to be of lower ability than those who would have been university graduates in both periods, that would affect the lower not the upper part of the university educated distribution.

In Charts 10 through 13, we present density plots for each year for specific birth cohorts, as defined earlier. Note that this contrasts with Charts 2 through 5, where we examine the same people of the same age in the two surveys. The Chart 2-5 comparisons thus present comparisons across cohorts (at the same age) while the Chart 10 through 13 comparisons follow individual cohorts through time.

**Chart 10**

**Document literacy, age 26 to 35 in 1994**

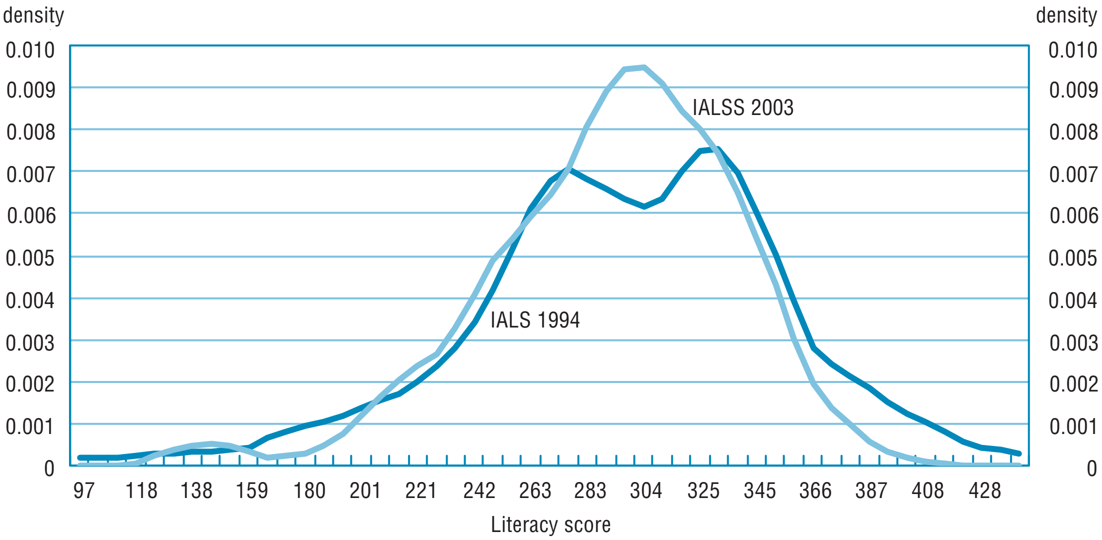



**Chart 11**

**Document literacy, age 36 to 45 in 1994**

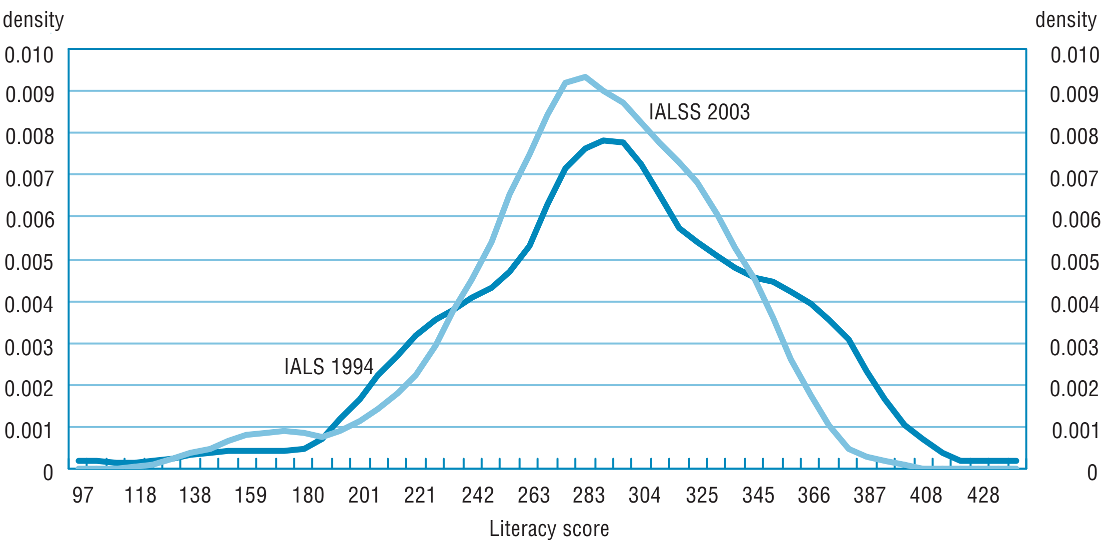


**Chart 12**

**Document literacy, age 46 to 55 in 1994**

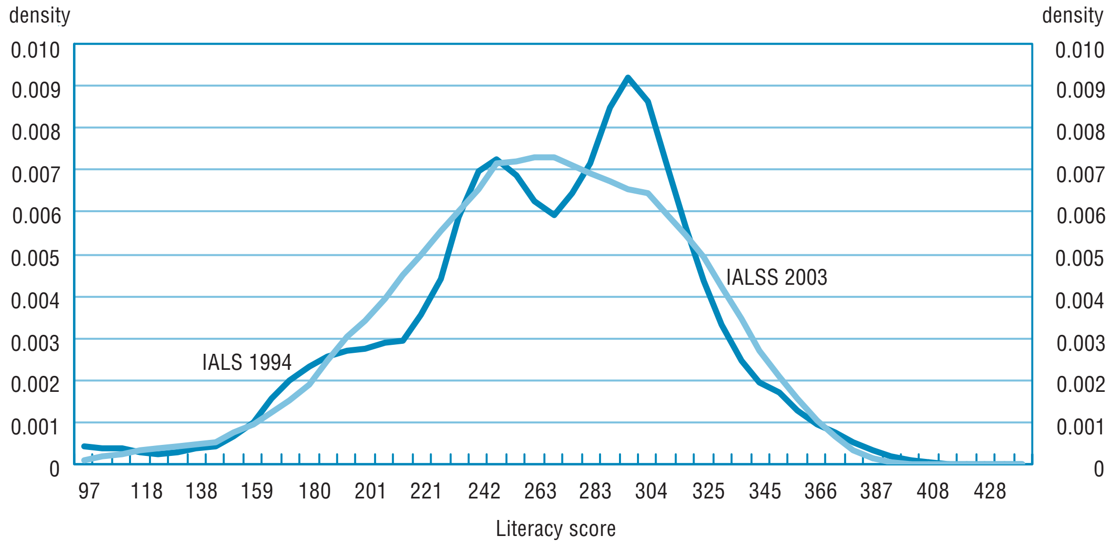



**Chart 13**

**Document literacy, age 56 to 65 in 1994**

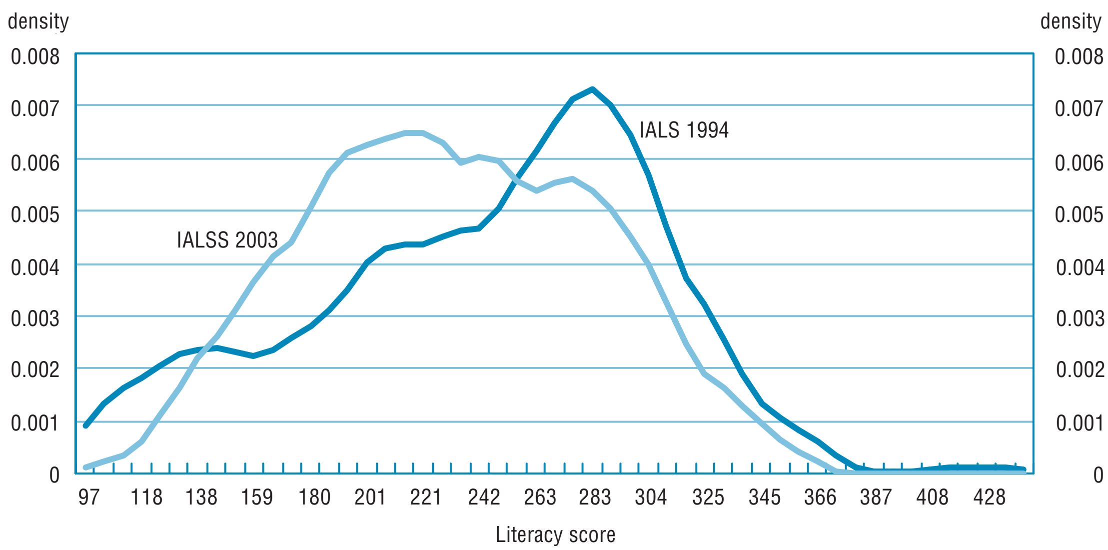


Chart 10 shows the plots for the cohort which was age 26 to 35 in 1994 and 35 to 44 in 2003. The plot indicates little difference at the low end of the distribution for the two years but a clear worsening at the top end. In fact, the 10th percentile of the literacy distribution increases from 223 to 232 for this cohort while the 90th percentile falls from 363 to 344. The cohort that was age 36 to 45 in 1994 shows a very similar pattern, plotted in Chart 11. In contrast, the densities for the cohort that was 46 to 55 show, if anything, an improvement from 1994 to 2003 (Chart 12). The 10th percentile for that cohort increased from 187 to 200 and the 90th percentile increased from 320 to 328 between the survey years. Again, this is important for considerations of the comparability of data since it shows that this cohort answers the full range of questions essentially as well (or slightly better) in 2003 as in 1994. The differences in patterns across cohorts would be very hard to rationalize simply as being due to more difficult top level questions being asked in 2003 than 1994. A more likely explanation of the patterns is that literacy skills deteriorate after leaving school but that the rate of deterioration declines with age.

Finally, in Charts 14 through 16 we plot densities following a particular education/ cohort group. Here, we define wider cohorts in order to ensure enough observations. Chart 14 shows the plots for those with less than a high school education. There is clear evidence of deterioration over time in literacy skills at all levels above about the 10th percentile. The same is also true for the high school graduates in Chart 15. The university graduates (shown in Chart 16) show mixed changes up to about the median but, again, clear deterioration above that point. Thus, there is evidence of deterioration of higher level literacy skills for all education groups.

**Chart 14**

**Document literacy, less than high school, age 26 to 45 in 1994**

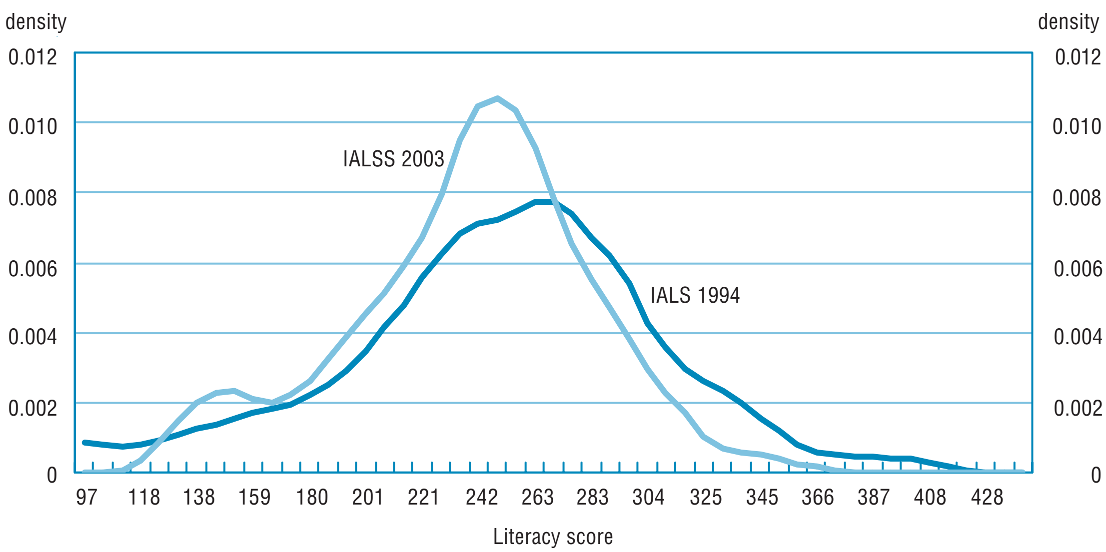


**Chart 15**

**Document literacy, high school, age 26 to 45 in 1994**

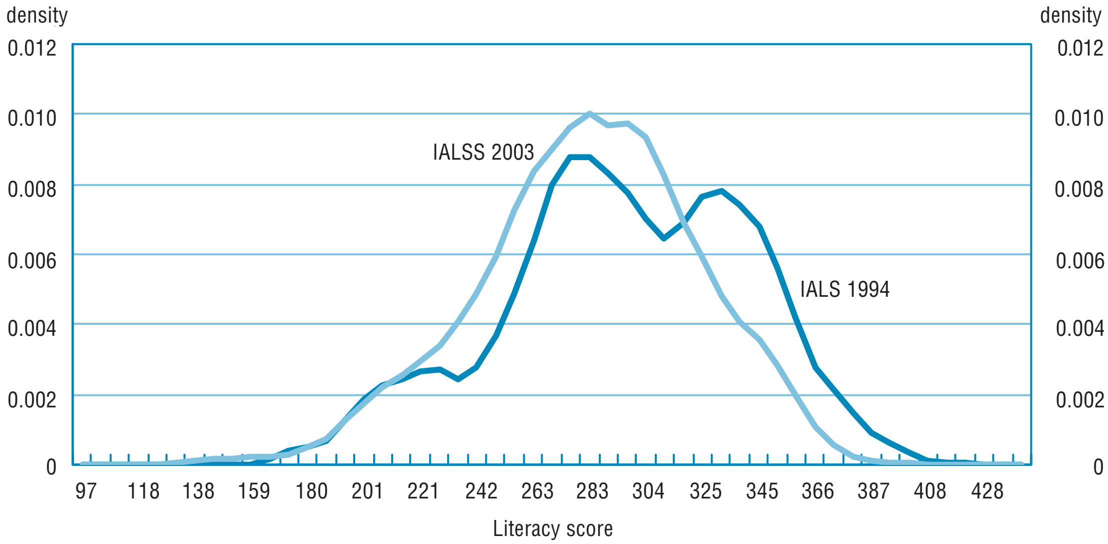
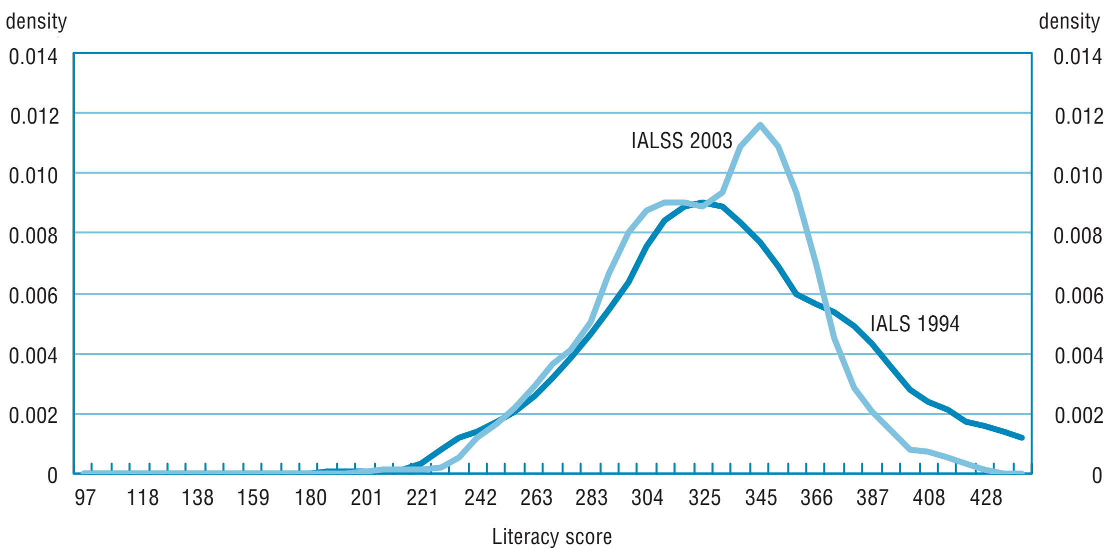



**Chart 16**

**Document literacy, university, age 26 to 45 in 1994**



To pursue these issues of ageing and cohort effects further, we estimate a series of regressions in which we pool the IALS 1994 and the IALS 2003 data and include cohort dummy variables in our standard specification. The public use version of IALS 1994 does not contain a continuous age variable so we adjust our standard specification by using age dummy variables instead of age and age squared variables. We report the results from this specification using simple OLS estimation in the OLS column of Table 4. Note that all of the following specifications include the parental education and immigrant status variables. In all cases, these variables exhibit the same patterns as in earlier estimation (with low parental education having a negative impact but the remaining variables not being significant) and we do not report them for the sake of brevity. As in the earlier specification, the schooling variables enter strongly and significantly. The schooling effect is larger than we witnessed in earlier tables. This is evident once we include the cohort dummy variables. Essentially, the schooling variable in the earlier specifications was partly picking up the cohort effects shown here: older cohorts have both less schooling and higher literacy, leading to an under-estimate of the true impact of schooling.

**Table 4**

**Pooled regressions including cohort effects**

| Variable | OLS | Standard error | 10th quantile | Standard error | Median | Standard error | 90th quantile | Standard error |
|---|---|---|---|---|---|---|---|---|
| Years school | 0.083*** | (0.0067) | 0.13** | (0.0052) | 0.086*** | (0.0031) | 0,043*** | (0.0028) |
| School squared | -0.002*** | (0.002) | -0.0033*** | (0.0002) | -0.0021*** | (0.0001) | -0.0009*** | (0.0001) |
| Female | -0.0087 | (0.0058) | -0.0064 | (0.0083) | -0.016 | (0.0049) | -0.0062* | (0.0036) |
| **Age group** | | | | | | | | |
| 36 to 45 | -0.024* | (0.013) | -0.0074 | (0.018) | -0.018 | (0.011) | -0.05*** | (0.0083) |
| 46 to 55 | -0.05** | (0.016) | 0.0034 | (0.025) | -0.048*** | (0.015) | -0.12*** | (0.011) |
| 56 to 65 | -0.09*** | (0.021) | -0.084** | (0.031) | -0.099*** | (0.019) | -0.17*** | (0.014) |
| 65 and over | -0.19*** | (0.033) | -0.18*** | (0.039) | -0.18*** | (0.026) | -0.3*** | (0.016) |
| Cohort 2 | 0.018* | (0.01) | -0.013 | (0.018) | 0.018* | (0.0097) | 0.055*** | (0.0078) |
| Cohort 3 | 0.032** | (0.016) | 0.031 | (0.025) | 0.042** | (0.016) | 0.082*** | (0.012) |
| Cohort 4 | 0.053** | (0.029) | 0.052* | (0.032) | 0.055** | (0.023) | 0.13*** | (0.014) |
| Cohort 5 | -0.01 | (0.033) | -0.053 | (0.039) | -0.004 | (0.025) | 0.11*** | (0.015) |
| Constant | 4.97*** | (0.054) | 4.39*** | (0.047) | 4.98*** | (0.025) | 5.48*** | (0.021) |
| Observations | 14,734 | ... | 14,734 | ... | 14,734 | ... | 14 734 | ... |

... not applicable
* Statistically significantly different from zero at the 10% level of significance.
** Statistically significantly different from zero at the 5% level of significance.
*** Statistically significantly different from zero at the 1% level of significance.

**Notes:** All specifications also include controls for parental education and immigration status (not reported).
Cohort 1 (the omitted category) consists of people who were age 26 to 35 in 1994, Cohort 2 consists of people who were 36 to 45 in 1994, Cohort 3 consists of people who were 46 to 55 in 1994, Cohort 4 consists of people who were 56 to 65 in 1994 and Cohort 5 consists of people who were over 65 in 1994.

The age and cohort group effects in the pooled OLS results show interesting patterns mirroring those seen in the charts. In particular, when we estimate our standard specification using IALS 2003 and replacing age and age squared with the age interval dummy variables, we again find a relatively flat age profile for much of the age range. In particular, the first two age dummy variables (corresponding to the 36 to 45 and 46 to 55 age groups) do not have large coefficients and do not show statistically significantly different effects relative to the base age group (the 26 to 35 age group). The older two age groups show negative effects, with the 65 and over age group dummy having a coefficient of -0.16. This fits closely with the estimated patterns using age and age squared variables. Once we include the cohort dummy variables, however, the age effects assume a continuous, downward sloping profile, with the 36 to 45 year olds having 2.4% lower average literacy and the 46 to 55 year olds having 5% lower average literacy than the base group. At the same time, all the cohort dummies apart from the last one have significant, positive effects that increase with the cohort. The cohort that was 56 to 65 in 1994, for example, has average literacy levels that are 5% higher than those for the base cohort, which was 26 to 35 in 1994. The very oldest cohort has literacy levels that are approximately the same as the youngest cohort. Overall, the implication from these results is that the flat literacy age profile (at least up to age 55) arises from a combination of declining literacy with age and lower average literacy for more recent cohorts.

We investigate these phenomena further using quantile regressions in the remaining columns of Table 4. Quantile regressions are run using the whole sample of individuals and show how, for example, the 10th percentile of the literacy distribution conditioning on the included covariates changes as each of the covariates change marginally. Thus, the coefficient on schooling can be used to answer the question, "how does the 10th percentile of the literacy score distribution for people with 11 years of schooling differ from the 10th percentile of the literacy score distribution for people with 10 years of schooling?" Several points stand out from these regressions. First, the impact of schooling declines across quantiles: years of school have a greater impact in shifting lower quantiles than upper quantiles. In examining this evidence, it is important to take account of the second order terms since the schooling effect declines faster with additional years of school at the lowest quantile while it is closer to linear at the top. This means that schooling has a much larger effect in shifting the bottom than the top of the distribution at low years of schooling but about equal effects at the top and bottom of the distribution around 16 years of schooling. Second, the declining literacy with age effect is much larger at the top than the bottom of the distribution. Thus, as in the charts, the bottom of the distribution for various age groups is relatively similar but older age groups have much lower 90th percentiles. According to the table, for example, the 56 to 65 year old group has a 90th percentile that is 17% lower than the 90th percentile for the base (26 to 35 year old) group, controlling for education and parental education. Third, while the cohort effects are evident across the distribution, they are much stronger at the top end. Thus, the cohort who was 56 to 65 in 1994 (i.e., who were born between 1929 and 1938 have a 90th percentile that is 13% higher than that for the base group (who were born between 1959 and 1968. The intriguing question is whether this reflects a decline in effectiveness of the school system. The fact that the charts show some improvements in literacy across cohorts at the bottom end of the distribution while the regression results show declines is likely due to the fact that the latter control for education and parental education: newer cohorts have more education, as do their parents. Taken together these results may suggest that schools are doing a poorer job of imparting literacy at any given level but that there have been real benefits to the fact that successive generations have attained greater schooling.

In Table 5, we present quantile regressions for those with a high school or less education and those with a university education in order to take a closer look at education related effects. For the high school or less educated group, the ageing effects are virtually non-existent at the lowest quantiles, which perhaps reflects that there is little loss with age of very basic literacy. Also, there is evidence that the lowest quantiles for this education group have actually increased across cohorts. At the top end of the distribution, though, the results are much like the overall results in Table 4: strong declines with age combined with declines across recent cohorts. Thus, it seems possible that changes to public and high school education have been equalizing across cohorts, with improved literacy for those at the bottom being offset by not providing as good literacy training for those at the top. For the university educated, the declines with age are evident across the whole distribution. There is also evidence of declines across cohorts but these are not statistically well defined. This could be simply a sample size problem or it could suggest that the real difficulties in literacy across generations are associated with other post-secondary education rather than with university education. We need to study this point further.

**Table 5**

**Pooled quantile regressions with cohort effects, by education group**

| | High school or less | | | | | |
|---|---|---|---|---|---|---|
| Variable | 10th quantile | Standard error | Median | Standard error | 90th quantile | Standard error |
| Female | 0.064*** | (0.013) | -0.011 | -(0.0049) | -0.0096** | (0.0047) |
| **Age group** | | | | | | |
| 36 to 45 | 0.025 | (0.021) | -0.037*** | (0.01) | -0.027** | (0.012) |
| 46 to 55 | 0.019 | (0.035) | ... | (0.011) | -0.11*** | (0.015) |
| 56 to 65 | -0.032 | (0.068) | -0.11*** | (0.016) | -0.12*** | (0.024) |
| 65 and over | -0.017 | (0.12) | -0.28*** | (0.024) | -0.24*** | (0.027) |
| Cohort 2 | 0.0053 | (0.013) | 0.038*** | (0.01) | 0.061*** | (0.0094) |
| Cohort 3 | -0.053 | (0.053) | 0.015 | (0.017) | 0.046** | (0.016) |
| Cohort 4 | -0.23*** | (0.068) | 0.019 | (0.021) | 0.044* | (0.025) |
| Cohort 5 | -0.34*** | (0.11) | -0.027 | (0.027) | 0.042** | (0.024) |
| Constant | 5.45*** | (0.027) | 5.76*** | (0.011) | 5.87*** | (0.014) |
| Observations | 9,209 | ... | 9,209 | ... | 9,209 | ... |
| | **University** | | | | | |
| Variable | 10th quantile | Standard error | Median | Standard error | 90th quantile | Standard error |
| Female | -0.021** | (0.01) | -0.0046 | (0.0069) | 0.006 | (0.0084) |
| **Age group** | | | | | | |
| 36 to 45 | -0.064** | (0.027) | 0.0075 | (0.012) | -0.09* | (0.05) |
| 46 to 55 | -0.074* | (0.042) | -0.021 | (0.015) | -0.12*** | (0.022) |
| 56 to 65 | -0.18*** | (0.045) | -0.092*** | (0.017) | -0.15*** | (0.026) |
| 65 and over | -0.31*** | (0.093) | -0.16*** | (0.032) | -0.15*** | (0.038) |
| Cohort 2 | 0.012 | (0.024) | -0.025 | (0.017) | 0.0071 | (0.009) |
| Cohort 3 | 0.043 | (0.032) | 0.015 | (0.021) | 0.018 | (0.018) |
| Cohort 4 | 0.085 | (0.085) | 0.038 | (0.039) | 0.011 | (0.023) |
| Cohort 5 | 0.12 | (0.11) | 0.0071 | (0.041) | -0.12*** | (0.035) |
| Constant | 5.73*** | (0.032) | 5.8*** | (0.017) | 5.99*** | (0.015) |
| Observations | 2,163 | ... | 2,163 | ... | 2,163 | ... |

... not applicable

* Statistically significantly different from zero at the 10% level of significance.

** Statistically significantly different from zero at the 5% level of significance.

*** Statistically significantly different from zero at the 1% level of significance.

**Notes:** All specifications also include controls for parental education and immigration status (not reported).

Cohort 1 (the omitted category) consists of people who were age 26 to 35 in 1994, Cohort 2 consists of people who were 36 to 45 in 1994, Cohort 3 consists of people who were 46 to 55 in 1994, Cohort 4 consists of people who were 56 to 65 in 1994 and Cohort 5 consists of people who were over 65 in 1994.

Before leaving the cross-generational/age discussions, it is important to provide a caveat. This discussion hinges on the test scores being comparable across the IALS 1994 and IALSS 2003. There is reason to believe that this is the case (since the two share questions and the remaining questions have been scaled to allow for such comparisons, and also because of the differences in patterns across groups discussed above), but it is still a potential source of concern. The results here would be strengthened if we had access to the scores just for the 40% of questions that are common across the two surveys.

## 4.3 Post-school, work related literacy acquisition

The results in the previous section indicate that adults lose literacy skills with years since leaving school. In this section, we investigate a "use it or lose it" model of literacy. In particular, we examine whether individuals who use literacy skills on their jobs maintain higher literacy, controlling for age. To do so, we re-run our basic specification using the IALSS 2003 including variables on literacy use at work and occupation. The literacy use at work questions ask about frequency of performing reading, writing and mathematical tasks. Thus, for reading, questions are asked about five tasks. There are also questions on five writing tasks and five math tasks. We construct dummy variables equalling 1 if the individual responded that he or she performed four or five of the reading related tasks at least once a week and similar variables for the writing and math tasks. We also constructed dummy variables corresponding to performing one to three of the tasks at least once a week for each of reading, writing and math. Finally, we constructed a dummy variable corresponding to individuals who answered that they performed all of the tasks in a given area (e.g., reading) "rarely". In the regressions, the omitted category is then people who responded that they used some of the tasks but at a frequency of less than once a week.

We present simple OLS results from the basic regression including these literacy-at-work variables plus a job tenure variable in the OLS 1 column of Table 6. In order to reduce clutter in the table, we control for but do not report the parental education and immigration. Their estimated effects are very similar to what is reported in the earlier tables. The sample includes all individuals, whether working or not. The non-workers are captured in the dummy variable corresponding to job tenure being unknown. The estimated effect of job tenure is positive and statistically significant but very small: one extra year of job tenure increases literacy by 0.06%. Thus, literacy effectively does not change with job tenure. The literacy at work variables indicate that those who use literacy intensively at work (who report performing 4 or more literacy tasks in a group at least once a week) have higher literacy. Performance of reading tasks and performance of writing tasks at work are highly correlated, with 83% of those who say they read rarely also reporting writing rarely at work. Thus, in interpreting the coefficients, it is best to add together the coefficients from the reading and writing scores. Based on this, a person who performs four or more reading and four or more writing tasks per week has approximately 3.7% higher literacy on average. At the other end, a person who rarely writes or reads at work has approximately 2% lower average literacy. These effects are large enough to be noteworthy, but are not huge. That is, there is clearly a correlation between literacy of the individual and literacy use at work, but those who use literacy skills at work are not substantially more literate. Note that we do not know which way any causality runs behind these results. Literate individuals may be more likely to be employed in high literacy jobs, or high literacy jobs may help individuals maintain their literacy, or some combination of the two. Math use is not as correlated with the other two types of literacy at work groups as they are with each other. Thus, of the people whom we classify as reading rarely, only 43% say they also use math rarely. The estimation results do not report a positive effect on document literacy from using math often at work but those who use it rarely have over 3% lower literacy than the base group.

**Table 6**

**Regressions including literacy use at work and occupation variables**

| Variables | OLS 1 | Standard error | OLS 2 | Standard error |
|---|---|---|---|---|
| School | 0.047*** | (0.0046) | 0.0478*** | (0.0046) |
| Schooling squared | -0.0009*** | (0.0002) | -0.001*** | (0.0002) |
| Female | -0.009** | (0.0046) | -0.01** | (0.0049) |
| Age | 0.0043*** | (0.0009) | 0.004*** | (0.0009) |
| Age squared | -0.0075*** | (0.001) | -0.0072*** | (0.001) |
| Tenure | 0.0006** | (0.0003) | 0.0005** | (0.0003) |
| Tenure unknown | -0.021 | (0.0159) | 0.025 | (0.026) |
| **Literacy use at work** | | | | |
| Read: At least once a week 4 or more categories | 0.024** | (0.011) | 0.016 | (0.011) |
| Read: At least once a week 3 or more categories | 0.019* | (0.011) | 0.015 | (0.011) |
| Rarely read | -0.0045 | (0.016) | -0.0047 | (0.016) |
| Write: At least once a week 4 or more categories | 0.013 | (0.0093) | 0.0088 | (0.0094) |
| Write: At least once a week 3 or more categories | 0.013 | (0.0077) | 0.0098 | (0.0078) |
| Rarely write | -0.023** | (0.01) | -0.021** | (0.01) |
| Math: At least once a week 4 or more categories | -0.01 | (0.010) | -0.011 | (0.010) |
| Math: At least once a week 3 or more categories | -0.0062 | (0.0097) | -0.0064 | (0.0098) |
| Rarely use math | -0.035** | (0.015) | -0.033** | (0.015) |
| **Occupations** | | | | |
| Armed forces | ... | ... | 0.044** | (0.016) |
| Professionals | ... | ... | 0.006 | (0.0077) |
| Technicians | ... | ... | -0.002 | (0.0073) |
| Clerks | ... | ... | 0.006 | (0.01) |
| Service workers | ... | ... | 0.02** | (0.009) |
| Skilled agriculture | ... | ... | -0.019 | (0.019) |
| Craft | ... | ... | -0.004 | (0.0093) |
| Plant and machine operators | ... | ... | -0.024** | (0.010) |
| Elementary occupations | ... | ... | -0.052** | (0.014) |
| Occupation missing | ... | ... | -0.067** | (0.027) |
| Constant | 5.16*** | (0.04) | 5.19*** | (0.041) |
| Observations | 14,527 | ... | 14,527 | ... |
| R-squared | 0.54 | ... | 0.54 | ... |

... not applicable

* Statistically significantly different from zero at the 10% level of significance.

** Statistically significantly different from zero at the 5% level of significance.

*** Statistically significantly different from zero at the 1% level of significance.

**Note:** All specifications also include controls for parental education and immigration status (not reported).

In the OLS 2 column of Table 6, we add in occupational dummy variables. Their inclusion has some effect on the literacy at work variables but mainly in increasing their standard errors due to collinearity rather than substantially reducing their estimated effects. Thus, reading and writing often at work still implies 2.5% higher literacy and doing so rarely implies 2.1% lower literacy. The base occupational group is composed of managers and relative to them, those in the professions, clerical and technical jobs all have similar literacy levels. Interestingly, so do skilled agricultural workers and craftsmen and trades workers (recall that these results hold after controlling for education). On the other hand, service workers, machine operators and elementary labourers all have significantly lower literacy. Again, though, these effects are not huge. For example, the difference in average literacy between a manager and an elementary labourer is 5%. Compare this to a difference of over 30% between a person with 12 years of education and one with 16 years of education.

As a final check on the importance of literacy use at work, we re-ran the pooled specifications but only for those in the professional, technician and clerical occupations on the assumption that these are high literacy use occupations. If being in such occupations allows one to maintain literacy then we would expect the estimated negative age profile to be less steep for this group. However, when we carry out this estimation, the estimated age coefficients are very similar to those presented for all individuals in the OLS column of Table 4. This may be because of slippage in the samples (i.e. the set of people who are in these occupations in 1994 need not be the same as those in 2003) but this is likely small relative to the strength of the result. Overall, the evidence on the impact of literacy use at work is mixed. There are clear correlations between literacy use at work and average literacy levels but these are not large relative to the impact of schooling. Also, being in a high literacy occupation does not seem to forestall the decline in literacy with age experienced by the individuals in this sample.

# 5. The impact of literacy on earnings

## 5.1 A simple theoretical framework

We now turn to examining the instrumental importance of literacy in terms of its impact on earnings. We saw in the previous section that literacy is primarily determined by schooling. This immediately raises the possibility that commonly estimated returns to schooling are in fact capturing the returns to literacy. To both represent this point and provide a framework for interpreting the earnings regressions that follow, we begin by presenting a simple hedonic model of earnings generation. The discussion that follows is based on Green and Riddell (2003).

Assume that each worker potentially possesses a range of skills and can possess each of them in varying amounts (including zero). To make the exposition simpler, we will couch our discussion in terms of three skills. Individual earnings are determined according to some function of the skills an individual possesses and puts into use, as follows:

$$1)\ E_i = f\ (G_i^1, G_i^2, G_i^3) + \varepsilon_i$$

where $E_i$ are earnings for individual i in our sample year, $G_i^k$ is the amount of skill k that person i sells in the market, and $\varepsilon_i$ is a disturbance term that is independent of the skills. We think of the disturbance term as capturing either individual idiosyncratic events that are independent of the skill levels or measurement error in earnings. We interpret the f(.) function as an earnings generation function derived ultimately from an overall production function that is separable in other (non-skill) inputs. Thus, by characterizing the f(.) function, we can learn about the importance of the various skills and how they interact in production. To help in focussing ideas, we will think of $G^1$ as cognitive skills of the type measured in literacy tests, $G^2$ as other (perhaps manual) skills that are not captured in such tests and might be acquired through work experience, and $G^3$ as non-cognitive characteristics such as persistence that might be partly acquired through schooling.

Based on 1), we can construct a set of skill price functions given by,

$$2)\ r_k = \frac{\partial f}{\partial G^k}\ (G_i^1, G_i^2, G_i^3), \quad k = 1,\ldots,3$$

Note that the prices can vary according to the complete bundle of attributes the individual sells. We are interested in characterizing this set of skill price functions. Once we have done that, we will know the relative importance of the various skills in production and also whether the different skills are complements or substitutes in production.

Characterizing either 1) or 2) would be a relatively straightforward exercise if we observed the skills, $G_i^k$. Typically, of course, we do not observe them. What we do observe is some of the inputs used in generating the skills. To see how they enter our framework, consider a set of production functions for generating the skills:

$$3)\ G_i^k = h_k\ (\ yrs_i\ ,\ \exp_i\ ,\ \theta_i\ )$$

where h indexes the attribute type, yrs corresponds to years of formal schooling, exp is years of experience in the work force and θ is a vector of innate abilities. Note that we differentiate between abilities (which are innate) and skills (which may be acquired and are directly useful in production). The vector of abilities, θ, may include both cognitive and non-cognitive elements. That is, non-cognitive abilities such as persistence could be useful in generating both non-cognitive and cognitive skills.

If we do not observe the $G_i^k$'s directly, we can obtain an estimating equation by substituting the equations given by 3) into 1). This then yields a quasi-reduced form specification for annual earnings given by,

$$4)\ E_i = g(\ \ yrs_i\ ,\ \exp_i\ ,\ \theta_i\ ) + \varepsilon_i$$

Thus, we are considering an hierarchical model in which covariates standardly used in wage regressions are inputs into skill production and these skills (plus an independent error term) completely determine wages.

Now, let us examine the partial derivatives of earnings with respect to each of the skill production inputs (e.g., schooling, experience or an element of the ability vector). The partial derivative associated with one of the inputs, x, can be expressed as,

$$5)\ \frac{\partial E}{\partial x} = \frac{\partial f}{\partial G^1} * \frac{\partial h_1}{\partial x} + \frac{\partial f}{\partial G^2} * \frac{\partial h_2}{\partial x} + \frac{\partial f}{\partial G^3} * \frac{\partial h_3}{\partial x}$$

where we suppress the i subscript for simplicity. Thus, if x corresponds to years of schooling, yrs, then equation 5) says that the observed effect of an additional year of schooling reflects the effects of an extra year of education on the production of each attribute times the price paid for that attribute. It is apparent from equation 5) that with measures only of earnings and observable inputs used in producing attributes, we cannot make any statements about skill production or how skills combine in production apart from statements that either a critical combination of the derivatives on the right hand side of 5) are zero (and, hence, $\delta E/\delta x = 0$) or some of them are not. If we have individual observations on a skill, e.g., $G^1$, however, we can potentially say much more.

With $G^1$ observed, our quasi-reduced form earnings function becomes:

$$6)\ E_i = g^*(G_i^1,\ yrs_i\ ,\ \exp_i\ ,\ \theta_i\ ) + \varepsilon_i$$

The derivative of this function with respect to $G^1$ corresponds to the attribute price function, $r_1$ - though, now we need to express the price as a function of $yrs_i$, $exp_i$, and $\theta_i$ :

$$7)\ r_1 = \chi(G_i^1,\ yrs_i\ ,\ \exp_i\ ,\ a_i\ )$$

With the price function given in 7), we cannot fully specify the interactions of $G^1$, $G^2$ and $G^3$ in production but we can learn more about them. In particular, the derivatives of $r_1$ with respect to the skill production input, x, is equal to (again suppressing the i subscript):

$$8)\quad \frac{\partial r_1}{\partial x} = \frac{\partial r_1}{\partial G^2} * \frac{\partial h_2}{\partial x} + \frac{\partial r_1}{\partial G^3} * \frac{\partial h_3}{\partial x}$$

Further, we can consider the derivative of g* in equation 6) with respect to x:

$$9)\quad \frac{\partial g^*}{\partial x} = \frac{\partial f}{\partial G^2} * \frac{\partial h_2}{\partial x} + \frac{\partial f}{\partial G^3} * \frac{\partial h_3}{\partial x}$$

With observed values for the derivatives, $\delta r_1/\delta x$ and $\delta g^*/\delta x$, we may be able to place restrictions on the f function. Thus, the fact that $\delta h_2/\delta x$ and $\delta h_3/\delta x$ appear in all these functions raises the possibility of putting restrictions on the production functions for the non-observed skills based on sign and significance patterns in the observed derivatives. With these restrictions in hand, we may further be able to place restrictions on the $\delta r_1/\delta G^k$ terms. In our framework, these latter terms reflect interactions in production of the non-observed skills with $G^1$.

We can also learn something about the production of $G^1$ from the differences between the derivatives, 5) and 9). These derivatives (e.g., the derivative of earnings with respect to schooling first not conditioning and then conditioning on $G^1$) differ by the term $\delta f/\delta G^1 * \delta h_1/\delta x$. Thus, the difference between these observed derivatives reflect the extent to which the coefficient on, for example, schooling in a standard earnings regression reflects the channel of added schooling generating added earnings through added cognitive skill creation. Given that we observe $G^1$ directly, we can go further and derive insights into the production of $G^1$ (as reflected in the $\delta h_1/\delta x$ terms) through direct estimation. That is essentially what we did in the previous section, where our main conclusions were that literacy is primarily produced through formal schooling and deteriorates with time after the person leaves school. The latter effect may be offset to some (likely relatively minor) extent by having a job that uses literacy skills.

Note that, as expressed in equation 7), the skill price facing an individual will be a function of the ability vector, $\theta_i$. In terms of our framework, this implies that the impacts of elements of the $\theta_i$ vector on $r_1$ can be written as in equation 8) with x replaced by the relevant element of $\theta_i$. That is, if we could observe $\theta_i$, we could learn more about the interactions of the various skills in production. In terms of empirical implementation, the fact that $r_1 = \delta E/\delta G^1$ could vary with unobservables points to the use of quantile regressions since they effectively allow us to observe derivatives of earnings with respect to observable variables at the different values of the unobservables that generate the various conditional quantiles. Moreover, if $\theta_i$ were a scalar rather than a vector, the h() functions were monotonic in $\theta_i$ and the f function were monotonically increasing in skills then increasing quantiles of the earnings distribution, conditional on $G^1$, yrs and exp, would be associated with increasing values of $\theta_i$ and we could sign $\delta r_1/\delta\theta$ based on how $\delta E/\delta G^1$ varies across increasing conditional earnings quantiles.

## 5.2 Estimation results

We present estimation results from mean regressions using the log of weekly wages in Table 7. We run the regressions on a restricted sample that includes only individuals with weekly wages that are over $50 and below $20,000 per week to eliminate a few extreme outliers. We also cut out the self-employed because their earnings will not fit in a model of the type outlined above without making further extensions to include returns to capital. The OLS 1 column shows the results from a standard regression with a female dummy, years of schooling, experience and experience squared as covariates. The results are extremely standard in terms of their magnitudes and sign patterns (see Card (1999) for a review of the very large relevant literature). In the OLS 2 column, we add our average literacy variable. This variable is the simple average of the 4 literacy scores. We take this approach because this is what is suggested by our factor analysis of the scores. Essentially, that analysis indicates the scores are highly collinear. We have estimated specifications in which we include all 4 scores separately. In those estimations, Document literacy enters statistically significantly with a coefficient of 0.0021, Numeracy enters statistically significantly with a coefficient of 0.0011 and Problem Solving and Prose literacy have smaller, not statistically significant and offsetting coefficients. Note that these significant separate effects essentially add up to the estimated coefficient on average literacy presented in Table 7. This suggests that Numeracy may have separate effects from the other three types of literacy and that its effects are smaller than whatever is being captured (primarily) in the document score.

**Table 7**

**Earnings regressions**

| | OLS 1 | Standard error | OLS 2 | Standard error | IV 1 | Standard error | IV 2 | Standard error |
|---|---|---|---|---|---|---|---|---|
| Female | -0.41*** | (0.024) | -0.4*** | (0.024) | -0.4*** | (0.024) | -0.4*** | (0.024) |
| Years of school | 0.087*** | (0.0042) | 0.069*** | (0.0047) | 0.07*** | (0.013) | 0.05*** | (0.017) |
| Experience | 0.067*** | (0.0035) | 0.067*** | (0.0035) | 0.068*** | (0.0041) | 0.068*** | (0.004) |
| Experience squared | -0.0011*** | (0.0001) | -0.0011*** | (0.0001) | -0.0012*** | (0.0001) | -0.0012*** | (0.0001) |
| Average literacy score | ... | ... | 0.0026*** | (0.0003) | ... | ... | 0.0032*** | (0.0007) |
| Constant | 4.8*** | (0.077) | 4.28*** | (0.10) | 5.04*** | (0.16) | 4.34*** | (0.10) |
| Observations | 7,768 | ... | 7,768 | ... | 7,768 | ... | 7,768 | ... |
| R-squared | 0.38 | ... | 0.4 | ... | 0.38 | ... | 0.39 | ... |

... not applicable

*** Statistically significantly different from zero at the 1% level of significance.

Adding average literacy to our standard earnings regression leads to a reduction in the derivative of log earnings with respect to education from 0.087 without the literacy variable to 0.069 when it is included. This is a reduction of about 20%, suggesting both that literacy skills play an important role in the returns to education and that education has a substantial impact on earnings over and above the impact related to production of literacy skills. This effect, however, is noticeably smaller than observed in the IALS 1994 (Green and Riddell (2003)). This may reflect a more flawed earnings measure in the earlier dataset.

In contrast to the effect on the schooling coefficient, the coefficients on the experience variables are unchanged when we introduce the literacy variable. This is a direct reflection of the fact that literacy generation is not related to age or experience in the cross-section. In terms of the framework set out above, experience does not enter the literacy skill production function and so the first term on the right hand side of equation 5 is zero. The implication is then that the derivative with respect to experience is the same whether or not we condition on literacy. In fact, however, we have seen that age does enter the literacy production function in a negative way but it is offset by cohort effects that are declining across recent generations. If the IALS 1994 had a comparable earnings measure and we could estimate a pooled cohort regression of the type we estimated for literacy in the previous section then we would see that literacy does have an impact on those estimated coefficients. In particular, a standard finding that earnings increase with age for a given cohort estimated without the inclusion of the literacy variable would be converted to an even steeper slope with the inclusion of the literacy variable.[9] The implication is that any increase in earnings with age must be due to an even more substantial increase in returns to non-literacy skills of a type that are acquired through working since literacy skills themselves are actually declining.

Finally, the direct impact of literacy skills on earnings is substantial. A 25 point increase in the average literacy score (the equivalent of about 1/2 of a standard deviation in the literacy score distribution) has an impact equivalent to an extra year of schooling.

As mentioned in our theoretical discussion, our estimation may be affected by standard omitted variables bias. In particular, the error term in the regression may include various types of ability which are correlated with the included variables. Typically, ability is assumed to affect both schooling choices and earnings, leading to biased estimates. Given our specified model, if we assume that unobserved cognitive abilities only affect the generation of cognitive skills and other, non-cognitive abilities do not affect the generation of cognitive skills then literacy will not be correlated with the error term and does not, itself, represent an endogeneity problem. However, we still need to address the potential endogeneity of schooling. We do this using province in which the individual resided and interactions of it with age as our instrument as we did in the literacy estimation. As before, we control for province of current of residence at the same time. In fact, all of the specifications include controls for current province whose effects are not reported in the table for the sake of brevity.

The results from instrumenting for schooling when the literacy variable is not included are reported in the IV 1 column of Table 7. Instrumenting reduces the coefficient on schooling slightly, which is suggestive of endogeneity problems with the simple OLS estimation. In the IV 2 column, we repeat this exercise but also include the average literacy effect. The instrumenting yields a somewhat smaller schooling coefficient than in the simple OLS results in the OLS 2 column. As in the simple OLS estimates, introducing the average literacy score reduces the schooling coefficient by approximately 20%. The literacy score coefficient itself rises somewhat but is of roughly the same order of magnitude as in the OLS 2 column. Thus, our main conclusions are not altered by instrumenting.

As discussed earlier, our theoretical framework points to advantages from using a quantile regression framework. We present the results of quantile regression estimation for the 10th, 25th, 50th, 75th and 90th quantiles in Table 8. The key implications from the estimation are as follows. First, returns to both schooling and experience decline across quantiles. The finding of heterogeneity in returns to education across the earnings distribution has been observed by previous authors. Buchinsky (1997) finds returns to education that rise across quantiles for all experience groups. Arias et al. (2001) estimate similar quantile regressions using US twins data and incorporating approaches to address endogeneity. With non-IV estimation, they find that the coefficient on education rises from the 10th to the 50th percentile but does not change across the upper portion of the distribution. When using instruments to address measurement error and twins status to address ability bias, their estimated schooling coefficients appear relatively similar across the distribution but are not very precisely estimated in the tails.

**Table 8**
**Quantile earnings regressions**

| | 10th quantile | Standard error | 25th quantile | Standard error | Median | Standard error | 75th quantile | Standard error | 90th quantile | Standard error |
|---|---|---|---|---|---|---|---|---|---|---|
| Female | -0.48*** | (0.048) | -0.46*** | (0.032) | -0.36*** | (0.020) | -0.36*** | (0.021) | -0.36*** | (0.032) |
| Years of schooling | 0.079*** | (0.011) | 0.065*** | (0.0065) | 0.069*** | (0.0041) | 0.067*** | (0.0042) | 0.057*** | (0.0069) |
| Experience | 0.088*** | (0.0063) | 0.084*** | (0.0041) | 0.059*** | (0.0025) | 0.05*** | (0.0024) | 0.042*** | (0.0036) |
| Experience squared | -0.0015*** | (0.0001) | -0.0015*** | (0.0001) | -0.001*** | (0.0001) | -0.0008*** | (0.0001) | -0.0006*** | (0.0001) |
| Average literacy score | 0.0022*** | (0.0007) | 0.0032*** | (0.0004) | 0.0029*** | (0.0003) | 0.003*** | (0.0003) | 0.0028*** | (0.0005) |
| Constant | 3.33*** | (0.19) | 3.67*** | (0.12) | 4.26*** | (0.083) | 4.62*** | (0.079) | 5.12*** | (0.13) |
| Observations | 7,768 | ... | 7,768 | ... | 7,768 | ... | 7,768 | ... | 7,768 | ... |

... not applicable
*** Statistically significantly different from zero at the 1% level of significance.

Perhaps the most interesting result in Table 8 is the relative lack of variation in the coefficient on the literacy score variable across the quantiles. While the coefficient for the 10th quantile appears substantively smaller than those at the other quantiles it is not actually statistically significantly so. Moreover, if we run quantile regressions at the quantiles directly surrounding the 10th quantile (e.g., the 5th and 15th quantiles), we obtain estimated literacy effects that are almost exactly the same as those reported for the upper quantiles – the 10th quantile appears to be a bit of a statistical outlier. In the context of our theoretical model this implies that cognitive skills do not interact with other attributes in earnings generation. Thus, other attributes such as beauty (Hamermesh and Biddle (1994)) and leadership skills (Kuhn and Weinberger (2005)) may contribute to individual earnings but their contributions are not enhanced by having more literacy skills. Literacy skills are not a silver bullet that both contributes directly to earnings and increases returns to other attributes. Put another way, to the extent that leadership skills contribute to earnings and perhaps productivity, these results do not suggest that firms should target leadership training by cognitive skill level: workers with various cognitive skill levels will receive the same benefit from the training.

Finally, our estimates of the effects of literacy presented here are identical to those we found with the IALS 1994 data (Green and Riddell (2003)). In both cases, with standard specifications, the coefficient on average literacy in a median earnings regression is 0.0029. We regard this as something of a fluke, since the earnings measures in the two studies are not the same and other variables have different estimated effects in the two studies. However, the key patterns estimated using the IALS 1994 hold up, including the size of the literacy effect and the fact that the literacy effect does not vary across quantiles. As in that study, we conclude that literacy is an important determinant of earnings but that there is a great deal of earnings variation that is accounted for by other factors.

# 6. Conclusion

In this paper, we examine literacy both from the perspective that it is an important entity in its own right and in terms of its impact on earnings. We find that literacy is considerably more equally distributed than income and that the level of literacy inequality has decreased between the time of the 1994 literacy survey and the more recent, 2003, survey. A closer examination reveals that there have been improvements in literacy at the low end of the literacy distribution but deterioration at the top end. The improvements partly reflect improved literacy among high school drop outs but appear mainly to result from improvements in the education level of the population. The declines at the top end are evident among high school graduates, university graduates and graduates of non-university post-secondary educated institutions. We find that successive birth cohorts have had poorer literacy outcomes at the top of the distribution, potentially pointing to an education system that is doing better for those at the low end but doing a poorer job of generating literacy for those at the top. We also find that literacy tends to decline with age after leaving formal schooling and that schooling itself is the prime driver of literacy. Parental education is also strongly related to literacy, but in an interesting way: having parents who are high school drop outs has a strong negative impact on literacy but there is almost no added impact from parental education rising above high school graduation. There is some evidence of a correlation of literacy with the use of literacy skills at work, suggesting (along with the declines with age) a "use it or lose it" model of post-schooling literacy, but these effects are not large.

In terms of earnings generation, literacy plays a substantial role. According to our Ordinary Least Squares estimates, a 25 point increase in the average literacy score (the equivalent of about 1/2 of a standard deviation in the literacy score distribution) has an impact equivalent to an extra year of schooling. When we introduce the literacy variable into a standard earnings regression, the coefficient on the schooling variable is reduced by approximately 20%, implying that about one-fifth of the typically measured impact of schooling on earnings arises because schooling generates higher levels of literacy. Our instrumental variables results seem to imply a more substantial role than this for literacy in accounting for returns to schooling but we need more investigation to be sure of this result. The fact that the coefficient on schooling remains strong and positive indicates, further, that schooling also affects earnings through its impact on non-cognitive traits such as leadership skills and reliable attendance at work.

# References

Aris, O., K.F. Hallock and W. Sosa-Escudero (2001). *"Individual Heterogeneity in the Returns to Schooling: Instrumental Variables Quantile Regression Using Twins Data,"* Empirical Economics (26), pp. 7-40.

Buchinsky, M. (1997). *"Recent Advances in Quantile Regression Models: A Practical Guideline for Empirical Research,"* The Journal of Human Resources, 33(1), pp. 88-126.

Card, D. (1999). *"The Causal Effect of Education on Earnings,"* in O. Ashenfelter and D. Card (eds.) Handbook of Labor Economics, Vol. IIIa. Amsterdam: Elsevier.

Crompton, S. (1996). *"The Marginally Literate Workforce,"* Perspectives on Labour and Income, Statistics Canada, Summer.

Frenette, M., D.A. Green and K. Milligan (2006). *"Revisiting recent trends in Canadian after-tax income inequality using Census data."* University of British Columbia and Statistics Canada.

Green, D.A. and Riddell, W.C. *"Literacy and Earnings: An Investigation of the Interaction of Cognitive and Unobserved Skills in Earnings Generation,"* Labour Economics 10 (April 2003) 165-84.

Hamermesh, D.S. and J.E. Biddle (1994). *"Beauty and the Labor Market,"* American Economic Review, 84 (No. 5) pp. 1174-94.

Kuhn, P. and C. Weinberger (2005). *"Leadership Skills and Labor Market Outcomes,"* Journal of Labor Economics, 23, July, 395-436.

Sen, A. (1999). Development as Freedom. New York: Anchor Books.

# Appendix A

## Education definitions in the IALS 1994, IALSS 2003, and LFS

In this appendix, we examine the different questions concerning educational attainment in the IALS 1994 and the IALSS 2003, comparing them to responses in the Labour Force Survey (LFS) in 1994 and 2003.

The IALS 1994 and IALSS 2003 have similar but not identical education questions. Both ask the respondent about the number of years they have gone to school, with zero years being an option. The IALSS 2003 then has a question asking whether the respondent completed high school, whether directly or by a high school equivalence test. IALS 1994 does not include this latter question. Both surveys include the question, "What is the highest level of schooling that you have ever completed?" However, they differ in the possible categories of response. The IALS 1994 has more possible categories at the sub-high school completion level but IALSS 2003 has more categories between the high school and university levels (5 compared to 1 in the IALS 1994). The latter difference raises the possibility of differences in how people who obtain trade certificates and apprenticeships are treated. These categories are explicitly listed in the IALSS 2003 question, giving even high school drop outs who went on to get other certificates the chance to record a higher level of education. With the shorter list in the IALS 1994, where the relevant category is the very general sounding "completed non-university Post-Secondary, one might imagine that respondents might neglect to mention apprenticeships, for example.

We have experimented with several education category definitions using the IALSS 2003 questions. We focus our attention on trying to match the IALS 1994 with the IALSS 2003 rather than the other way around because the education questions and categories are more extensive in the IALSS 2003. The choice of our categorization has little impact on the results we present in the body of the paper. We settled on a categorization in which we classify individuals who respond that they have not completed high school (to the direct question on that point) and state that their highest level of schooling is either less than high school, high school diploma, or some post-secondary education that did not result in a certificate or degree as having less than high school education. Our assumption in this is that these non-high school graduates would not have responded that their highest level of schooling was "completed non-university Post-Secondary" in the 1994 questionnaire. We classify individuals who give the same responses to the highest level completed question but state they have graduated from high school in response to the earlier question as high school graduates. Our college category includes all categories in the highest level of schooling question that are above the not-completed post-secondary category but below the bachelor's degree category. We gather all university degrees at the BA level and above together in the university category.

The resulting categorization yields the following comparison between the two datasets, where each entry in the IALS 1994 and IALSS 2003 columns shows the proportion in the corresponding category in the given year.

**Table A.1**

**Education category comparisons: IALS and IALSS**

| | IALS 1994 | IALSS 2003 | Difference |
|---|---|---|---|
| | | percent | |
| Less than high school | 36.6 | 25 | -11.6 |
| High school | 30.1 | 28 | -2.1 |
| Non-university post-secondary | 16.8 | 26.2 | 9.4 |
| University | 16.5 | 20.9 | 4.4 |

The results show a strong decline in the lowest education category and strong increases in the two post-secondary categories.

As a point of comparison, we also formed the proportion in corresponding categories from the Labour Force Survey for 1994 and 2003 as follows:

**Table A.2**

**Education category comparisons: LFS**

| | LFS 1994 | LFS 2003 | Difference |
|---|---|---|---|
| | | percent | |
| Less than high school | 31.8 | 22.2 | -9.6 |
| High school | 26.2 | 26.4 | 0.2 |
| Non-university post-secondary | 26.9 | 31.3 | 4.4 |
| University | 15 | 20.2 | 5.2 |

The broad patterns are similar in terms of the declines in the proportion who have less than a high school degree and the increases in the top two categories. The proportions in the university categories are also very similar across the two types of datasets in each year. However, the LFS has much higher proportions in the some post-secondary category, particularly in the first period, and correspondingly lower proportions in the bottom two categories. From this, we draw two conclusions. First, it is not clear which type of dataset (the literacy surveys or the LFS) should be viewed as more reliable. It is worth noting, in particular, that the LFS includes responses about the target individual's education by other household members if the target individual cannot be contacted. It is not known how many observations are affected by this "proxy response" and the potential measurement errors it introduces. In the literacy surveys, on the other hand proxy response is not an issue since the target individual must necessarily be contacted directly in order for them to take the test. On the other hand, this may itself raise issues of representativeness relative to the LFS since the LFS at least has some information on people who are hard to contact directly. Second, the broad trends are similar across datasets and thus the IALS datasets seem likely to provide an accurate picture of the impact of changes in the educational composition of the population over time.

# Appendix B

## First stage results for years of schooling

**Table B.1**
**Log literacy regressions**

| Variable | OLS | Standard error |
|---|---|---|
| Age | 0.26*** | (0.015) |
| Age squared | -0.27*** | (0.014) |
| **Mother's education** | | |
| Less than high school | -1.01*** | (0.13) |
| Some post secondary | 0.54*** | (0.14) |
| BA or more | 0.47 * | (0.26) |
| None reported | -1.56*** | (0.20) |
| **Father's education** | | |
| Less than high school | -0.77*** | (0.13) |
| Some post secondary | 0.5*** | (0.18) |
| BA or more | 1.31*** | (0.19) |
| None reported | -1.59*** | (0.19) |
| Immigrant mother | 0.44*** | (0.17) |
| Immigrant father | 0.49*** | (0.16) |
| Constant | 8.68*** | (0.40) |
| Observations | 13,868 | ... |
| R-squared | 0.23 | ... |

... not applicable
* Statistically significant at 10% level.
*** Statistically significant at 1% level.
**Note:** Specification includes a complete set of current province of residence dummy variables, a complete set of province during high school dummy variables, and a complete set of interactions of the latter with age.

# Endnotes

1. The other countries participating in this first round of the IALSS 2003 were Bermuda, Italy, Mexico (specifically, the state of Nuevo Leon), Norway, Switzerland and the U.S. The earlier IALS survey was carried out in over 20 countries during the period 1994 to 1998.
2. We truncated the density plot at 96 to focus attention on the main part of the distribution. This eliminates only five observations but because of the smoothing inherent in the kernel density plots, their inclusion would result in a plot that is visually dominated by a thin, long left tail. The plot uses Strata 7's default values for the kernel and smoothing parameters.
3. This number is obtained by taking the first derivative of the age profile and evaluating it at age 30. For this quadratic profile, this first derivative equals the coefficient on the linear age variables plus (two times the coefficient on the quadratic age variable times 30).
4. This calculation is made in the same way as the one described in footnote 3.
5. Note that the number of observations is now fewer because we have dropped observations for which province where the individual went to high school is not stated.
6. The test statistic corresponding to the joint hypothesis that the province of residence dummies and their interactions with age are significantly different from zero is distributed as F(22,13821) and has an associated p-value of 0.009.
7. This conclusion would not be true if province of residence when in high school has a direct impact on literacy outcomes. This might be the case if individuals living in resource rich provinces while in school purposefully under-invest in literacy because they know they are going to drop-out and take a job that does not require literacy when they are old enough. Essentially, we must assume that such a relationship does not exist, at least after controlling for current province.
8. Experience is defined according to the standard Mincer expression (i.e., experience = age - years of schooling - 6). The coefficient on the experience variable in our simplest OLS regression (which includes only a female dummy, schooling, schooling squared, experience and experience squared) is 0.00051 with a standard error of 0.00045 and the coefficient on the experience squared term is -0.000073 with a standard error of 0.0000083.
9. We cannot prove this result but it is strongly suggested by the relationship between literacy and age combined with standard omitted variable bias derivations.