SMARTEN UP: It's time to build essential skills

APPENDIX

(How the data was derived)

METHODS EMPLOYED TO LINK THE PROGRAMME FOR THE INTERNATIONAL ASSESSMENT OF ADULT COMPETENCIES (PIAAC) AND THE NATIONAL HOUSEHOLD SURVEY (NHS) DATASETS

Exploring the relationships between literacy, demographic and labour market characteristics required the imputation of literacy scores and literacy market segments onto the NHS file. The methods employed reproduce the PIAAC skill distribution exactly.

This Annex provides an overview of the methods that were used to impute literacy scores and literacy market segments on to the NHS file, how literacy demand levels were derived and how remedial costs and estimated benefits of eliminating literacy skill shortages were estimated and how rates of return were estimated.

The overall goal of the analysis was to impute a literacy score for each individual on the NHS individual file for 2011 for respondents aged 16 to 65. The imputations were based on a selection of personal characteristics that are associated with literacy scores such as age, gender, education, city size, immigration status, province, type of employment and occupation.

Using the relationships revealed in the PIAAC analysis determined the best estimate of an individual's literacy score and the chances that they would be at prose literacy levels 1, 2, 3, 4 or 5.

The analysis relied on individual records from two databases:

- → The PIAAC for 2011.
- → The National Household Survey for 2011 individual micro data files for Canada.

Analysis of the PIAAC data

The PIAAC data were used to perform a regression of reading literacy level on predictor variables. The regression was done for those individuals who had valid responses for all the variables of interest.

The dependent variable was the average of the 5 estimates of prose literacy provided by the PIAAC file. The results of these regressions gave regression coefficients that were subsequently used to predict the likely literacy levels of individuals on the NHS.

Independent variables that previous analysis had shown to be important predictors of literacy skill (Desjardins, 2004) were selected. Additionally independent variables had to be available on both the PIAAC and NHS and had to be code-able in a consistent fashion.

The coding of variables in PIAAC was changed from previous runs so that it could be consistent with NHS coding.

The regression coefficients are presented in the attached table. There were 27,287 observations in the regression and the resultant R2 was 23%.

- → The National Household Survey for 2011 individual micro data files for Canada.

FIGURE 1

REGRESSION ANALYSIS OF AVERAGE PROSE LITERACY: COEFFICIENT FOR EACH VARIABLE COMPARED TO REFERENCE

Regression to predict literacy value PIAAC for the population 16 to 65		
	Intercept	271.80
	Female	-5.07
Age group	16 to 25	25.40
	26 to 35	14.07
	36 to 45	8.71
	46 to 55	1.10
	56 to 65	–
Education	Less than high school	-65.49
	High school graduate	-35.80
	College diploma	-24.77
	Degree	–
Labour force status	Employed	-4.10
	Unemployed	-1.96
	Not in the Labour Force	–
Mother Tongue	English	14.12
	French	9.12
	Other	–
	Aboriginal	-12.76
	Immigrant	-23.14
Province	Newfoundland and Labrador	3.68
	Prince Edward Island	9.55
	Nova Scotia	8.45
	New Brunswick	6.40
	Quebec	8.65
	Ontario	15.69
	Manitoba	16.77
	Saskatchewan	5.76
	Alberta	15.66
	British Columbia	16.70
	Northwest Territories	–
Occupations	Not employed	8.41
	Management occupations	21.24
	Business, finance and administration occupations	19.54
	Natural and applied sciences and related occupations	25.95
	Health occupations	11.88
	Occupations in social science, education, government service and religion	16.21
	Occupations in art, culture, recreation and sport	25.04
	Sales and service occupations	5.51
	Trades, transport and equipment operations and related occupations	-3.54
	Occupations unique to primary industry	4.48
	Occupations unique to processing, manufacturing and utilities	–

Source: PIAAC, 2011

These regression coefficients are used for imputing literacy scores on the 500 point IALSS prose literacy scale on to micro data records from the 2005 to 2009 SLID files.

Imputation of reading literacy scores

Using NHS micro data files the best estimate of reading literacy score was determined for each individual based on their individual characteristics; age, gender, education, city size, immigration status, province, type of employment and occupation.

After generating this best estimate and the adjustment was made for the local literacy level as described above, individual score values are generated by simulating possible values using a normal distribution with mean equal to the best estimate and using a variance based on the Mean Squared Error of prediction. As well, a set of 10 possible literacy values were generated for each individual so one could determine the probability that they were at level 1, 2, 3, 4 or 5.

After imputation the imputed distributions of prose literacy and the proportions of the population at various literacy levels was compared to the PIAAC results. The following chart and associated table reveals that the distribution of average literacy scores by occupation from the two sources are in close agreement. The higher proportion of adults with literacy and numeracy skills at Level 1 is attributable to the fact that the NHS estimates include Indian Reserves.

The imputation procedure generated:

- An imputed Literacy Level which is one of 1 to 5.
- A Literacy Score roughly between 100 and 500 which is consistent with the level.
- An estimate of the confidence level associated with these imputed values.

Two forms of regression are used to generate the imputed values:

- Logistic regression where the dependent variable is the Literacy Level (1 to 5).
- Ordinary Least Squares regression where the dependent variable is the Literacy Score.

To assist in the imputation, the IALSS data were used to compare the percentiles of actual scores to percentiles of predicted Literacy Values.

Using the PIAAC survey the actual Literacy Scores are compared to predicted values (based on the OLS Regression).

This is done within each Literacy Level so one can compare the percentiles of the actual scores associated with the percentiles of the predicted values.

FIGURE 2

A COMPARISON OF PIAAC PUMF, PIAAC SHARE FILE AND NHS IMPUTED SKILL DISTRIBUTIONS

		Results from share/ COOL* version of PIACC				National Household Survey imputed values			Difference COOL* vs National household Survey		
	PIAAC PUMF	Literacy	Numeracy	Problem solving		Literacy	Numeracy	Problem solving	Literacy	Numeracy	Problem solving
Proficiency level		Percent			Proficiency level			Percent			
Level 0				0.1809							
Level 1	0.1289	0.163	0.2213	0.3706	**Level 1**	0.15	0.17	0.20	0.01	0.05	0.17
Level 2	0.3165	0.3196	0.3255	0.3642	**Level 2**	0.30	0.34	0.36	0.02	-0.01	0.00
Level 3	0.3998	0.3758	0.3196	0.0825	**Level 3**	0.39	0.34	0.35	-0.01	-0.02	-0.27
Level 4	0.1443	0.1215	0.1168	0.0015	**Level 4**	0.14	0.12	0.06	-0.02	-0.01	-0.06
Level 5	0.0106	0.0101	0.0156		**Level 5**	0.02	0.02	0.03	-0.01	-0.01	

Source: PIAAC, 2011, NHS, 2011.

* Statistics Canada Research Data Centre at the University of Ottawa

The imputation procedures for each individual on the NHS micro data files are as follows:

A Literacy Level (1 to 5) is imputed based on the Logistic Regression Coefficients. The imputed value is random using not only the coefficients but also the variance/ covariance matrix.

A preliminary Literacy Score is imputed based on the OLS Regression. This score may not be in the appropriate range for the imputed Literacy Level.

This preliminary Literacy Score is converted into a final score as follows:

- The preliminary Literacy Score is converted into a percentile of predicted scores (based on the PIAAC analysis) within the imputed Literacy Level.
- This percentile level is used to pick an actual score from the IALSS at this percentile level, within the Literacy Level.

This actual score is the imputed Literacy Score.

This imputation is repeated 10 times so that a variance in the various Literacy Scores and levels can be estimated. Imputation was based on Regressions with the following Covariates: Gender Education Level: 5 Categories Age Group Province Employed Full-Year (Yes / No) Occupation; 10 Categories CMA and Immigrant Status: Non-CMA, CMA – Non-Immigrant, CMA – Immigrant such as age, gender, education, mother tongue, city size, immigration status, province, type of employment and occupation.

CanadaWest FOUNDATION

CANADA WEST FOUNDATION
900–105 12th Avenue SE
Calgary AB T2G 1A1
www.cwf.ca