Scaling TOWES and Linking to IALS

Kentaro Yamamoto and Irwin Kirsch
March, 2002

In 2000, the Organization for Economic Cooperation and Development (OECD) along with Statistics Canada released *Literacy in the Information Age: Final Report of the International Adult Literacy Survey (2000)*. This assessment, conducted in three stages between 1994 and 1998, examined the proficiencies of the adult population in 23 different countries, including the United States and Canada. Up until this time comparable information on the literacy proficiencies of adult populations was quite limited with most of the information coming from either self reports of the highest level of completed education or international surveys of school-aged populations. The IALS assessment measured respondents' proficiencies along three literacy scales: prose, document and quantitative. Each scale was constructed to range from 0 to 500 and was divided into five levels that captured the increasing complexity and difficulty of the literacy tasks.

In March of 2000, we responded to a request for a field study design that would allow the development of the TOWES (Test of Workplace Essential Skills) to move forward toward an operational test that both reported results on three scales (Reading Texts, Using Documents and Quantitative Literacy) and that were linked to the International Adult Literacy Study. For some time, workplace-training practitioners had expressed the need for a test of essential workplace skills that would yield results similar to those reported by IALS. TOWES was designed to test generic literacy skills using materials found in Canadian workplaces and tasks typical of those carried out by Canadian workers. It is strongly tied to research conducted in Canada that focused on reading and essential skills in the workplace. This work, carried out under the Essential Skills Research Project (ESRP), was initiated by Human Resource Development Canada (HRDC). TOWES consists of a bank of items that were developed and field tested using authentic workplace documents and recreating actual tasks that are both similar to what individuals do at work and to what was measured in IALS. This report describes the models and procedures used to scale the TOWES field study results and link them to the IALS scales.

TOWES FIELD STUDY DESIGN

The TOWES field study collected information on 2,689 respondents through a background questionnaire and a series of assessment booklets containing prose, document, and quantitative literacy tasks. To achieve good content coverage for each of three literacy domains and to link TOWES to IALS, the number of tasks in the assessment had to be quite large. Altogether, 412 cognitive tasks were administered in the linking study. Yet, the time burden for each respondent also needed to be within an acceptable range. To accommodate these two conflicting requirements – in other words, to reduce respondent burden without sacrificing good representation of the content domain – each respondent was administered only a fraction of the pool of tasks, using a variant of matrix sampling. The design proposed for the field study is shown here.

Booklet	T Block	IALS	T Block	IALS	Number of New Cases
1	A1	IALS 1	B2	IALS 2	200
2	B3	IALS 3	C1	IALS 4	200
3	C2	IALS 5	D3	IALS 6	200
4	D1	IALS 7	A2	IALS 1	200
5	A3	IALS 2	B1	IALS 3	200
6	B2	IALS 4	C3	IALS 5	200
7	C1	IALS 6	D2	IALS 7	200
8	D3	IALS 1	A1	IALS 2	200

Booklet	Block 1	Block 2	Block 3	Number of New Cases
9	A2	B1	C3	150
10	B2	C1	D3	150
11	C2	D1	A3	150
12	D2	A1	B3	150
13	A3	B3	C3	150
14	C2	D2	A2	150
15	D1	B1	C1	150

Each respondent received one of the 15 test booklets containing either 3 blocks of the new TOWES literacy tasks or a combination of the TOWES and IALS tasks. Each block contained tasks from the three literacy domains. In total, some seven IALS blocks were selected to link the TOWES items and place them on the IALS literacy scales.

Since the item parameters for IALS are already established, the additional 400 cases for each IALS item (600 for blocks 1 and 3) were used to test these parameters against a new population of adults and to link the TOWES tasks to the prose, document and quantitative literacy scales. While this was not a balanced design, each of the TOWES items was administered to at least 500 respondents. In addition, the field study design assumes that these 15 booklets were administered to a random sample of adults covering a broad range of the ability distribution and who were representative to the extent possible with respect to education, gender and race/ethnicity. Moreover, the design assumes each of the 15 booklets was administered to a randomly equivalent sample of adults.

OVERVIEW OF THE ANALASES

This section of the report identifies several issues and summarizes the procedures that were followed in order to deal adequately with them. Included in the discussion are: scoring reliability, treatment of missing data, block-order effects, and estimating item parameters and linking TOWES to the IALS scales. A brief introduction is provided about each of the four issues and why they are important to the goals of this study before discussing each issue in detail in terms of the analytic procedures that were followed and the outcomes that were observed.

Scoring reliability and missing data

Since IALS and TOWES use open-ended tasks to measure literacy skills, the possibility exists that different scorers could apply the rubrics for each task in different ways. This would result in an item measuring different things depending on who was scoring it. High inter-rater agreement is a necessary first step to minimize the size of measurement error coming from scoring. It is important to note that any systematic errors introduced by scorers will result in bias and will be transferred to the estimation of item parameters.

Since the literacy tasks are open ended, training scorers and checking for consistency in applying the scoring rubrics are important steps in the scaling and linking procedures. Care is taken in developing the scoring rubrics and in the training of scorers to apply these rubrics consistently. Beyond scorer training, a procedure was set up to monitor scoring accuracy by evaluating a subset of scored responses for each item and each scorer. At the beginning of the scoring activities, almost all responses are monitored to identify problems with existing scoring guides or individual scorers who may not be applying one or more scoring rubrics in the same way as the other scorers. In addition, some precautions need to be made to ensure independence of the first and the second scores. For example, the two scores must be obtained from different scorers, and the second scorer should not be able to see the scores given by the first scorer.

Looking at the proportion of agreement is one way to assess scoring accuracy. Although Cohen's kappa is an alternative statistic, both methods produce essentially the same results. Hence, the proportion of agreement was used to assess the scoring accuracy for TOWES.

On average 200 responses were rescored for every item. It should be noted that all items regardless of whether they were retained or dropped for some reason were included in the evaluation of scoring reliability. The following table represents the summary results. It is clear that proportion of agreement averaged across each set of items is quite high and is comparable to what was found in the original IALS scoring study.

Scoring Reliability

	No. of items	**Reliability (proportion agreement)**
TOWES	305	.97
IALS	107	.96
TOTAL	412	.96

In addition to issues of scoring, all tests have respondents who choose to omit particular items or who may not have the opportunity to respond to an item because of time constraints or other circumstances. This problem can be magnified in assessment designs such as the one used in this study in which each respondent is given only a subset of the total item pool. As a result, statistical procedures can be implemented which minimize the bias that would be introduced if one assumed that not reached items are associated with ability. As a result, we distinguish between omitted and not reached items such that not reached items are assumed to be unrelated to ability and treated differently than omitted items.

The treatment of missing responses may introduce errors into the scaling procedure due to misattribution of the causes of not responding to items. In this study, as in IALS, all responses appearing before the last legitimate response made by the test taker were treated as omitted, while all responses occurring after the last response were treated as not reached. This results in omitted items being treated as wrong since a respondent had the chance to produce a response and chose not to and not reached items being treated as if they were not administered. From a scaling perspective, not reached items are not used in estimating proficiency while omitted items are a part of the estimation.

Block position effect

Because we are using a variant of matrix sampling in this study and each block of items appears in more than one position, it is important to determine whether the performance of an item is associated with its position in the assessment design. Position effects, if they are present, increase the size of measurement error and the extent to which the overall IRT model is well estimated. This could impact both how well the parameters for the new items are estimated as well as how stable the existing parameters are for the IALS tasks. If a significant position effect is detected it must be taken into account and a new model estimated.

A block position effect or testing for the interaction of item position with performance was evaluated. The field study was designed so that we could test for such a possibility. While item position was found not to be a factor in IALS it is important to test for such an effect due to the impact it can have on model fit and error estimation. Because of the nature of the design, it was most efficient to test for order effects for TOWES items that appeared in the first and third positions and each of the IALS items that appeared in the second and fourth positions. The table below shows the block level proportion correct averaged over blocks. Average proportion correct was slightly lower when blocks appeared later in booklets and it was more pronounced for IALS blocks. We used analysis of variance to evaluate the significance of this effect. We found the F statistic was not significant for both the TOWES and IALS blocks. Thus, there is no evidence for the block position effect.

Proportion correct of an item was calculated as follows,

$$P = \frac{\sum correct}{\sum correct + \sum wrong + \sum omit}$$

Note: The denominator represents the total attempts. Not-reached responses, consecutively missing responses at the end of a block, were not included in calculating proportions correct. Exclusion of not-reached items from proportions correct is consistent with scaling procedures applied to produce proficiency scores based on the IRT model for IALS.

Summary of Item Order Effects on Performance in TOWES and IALS

Block Order	TOWES blocks	IALS blocks
First/Second	68.5	75.8
Third/Fourth	66.5	71.8
Position effect	F= 2.96 (p=0.10) NS	

The scaling model

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

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

where

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

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

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

b_i is its locator parameter, characterizing its difficulty.

The main assumption of IRT is conditional independence. In other words, item response probabilities depend only on θ (a measure of proficiency) and the specified item parameters, and not on any demographic characteristics of examinees, or on any other items presented together in a test, or on the survey administration conditions. This enables us to formulate the following joint probability of a particular response pattern *x* across a set of *n* items.

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

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

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

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

Note that this is a monotone increasing function with respect to θ; that is, the conditional probability of a correct response increases as the value of θ increases. In addition, a linear

indeterminacy exists with respect to the values of θ_j, a_i, and b_i for a scale defined under the two-parameter model. In other words, for an arbitrary linear transformation of θ say $\theta^* = M\theta + X$, the corresponding transformations $a^*_i = a_i/M$ and $b^*_i = Mb_i + X$ give:

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

Since the scaling procedure retained the functional relationship between proficiency and conditional probability of correct responses, the identical transformation constants used for the IALS can be applied.

Scale linking and Item parameter estimation

There are two general ways to link the TOWES and IALS data. One involves estimating the ability distributions separately for each set of items and then matching the distributions. The second and more desirable way is to link at the item and scale level. This is preferred method because it maintains the underlying dimensionality of the scale or scales and was the method used in this study. The first step was to evaluate the IALS items by treating the existing item parameters as fixed. To the extent that the IALS item parameters are stable as estimated by various statistical procedures, we can then estimate parameters for the new items having them on the existing IALS scales.

In IALS, the estimation of item-parameters for the 2PL model has been carried out using a modified version of Mislevy and Bock's (1982) BILOG program. BILOG procedures are based on an extension of the marginal-maximum-likelihood approach described by Bock and Aitkin (1981). The IALS version of BILOG maximizes the likelihood

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

In the equation, $P(\underline{x}_{j,g}|\theta,\beta)$ is the conditional probability of observing a response vector $\underline{x}_{jg}$ of person j from group g, given proficiency θ and vector of item parameters $\beta = (a_1, b_1, \ldots a_i, b_i)$ and $f_g(\theta)$ is a population density for θ in group g. Prior distributions on item parameters can be specified and used to obtain Bayes-model estimates of these parameters (Mislevy, 1984). The proficiency densities can be assumed known and held fixed during item parameter estimation or estimated concurrently with item parameters.

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

the multinomial probabilities at the corresponding points that approximate $f_g(\theta)$ at $\theta = X_k$. If the data are from a single population with an assumed normal distribution, Gauss-Hermite quadrature procedures provide an "optimal" set of points and weights to best approximate the integral for a broad class of smooth functions. For more general f or for data from multiple populations with known densities, other sets of points (e.g., equally spaced points) can be substituted and the values of $A_g(X_k)$ may be chosen to be the normalized density at point X_k (i.e., $A_g(X_k) = f_g(X_k) / \sum_k f_g(X_k)$).

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

Common-item approaches can be used when the scales to be linked are based on item pools that share a set of items. IALS items are adapted verbatim to the TOWES scaling and linking study, which means that responses to the IALS cognitive items by the TOWES sample support the linkage between the TOWES scale and the IALS scale. This method is the most straightforward to link multiple scales.

The variant of concurrent calibration method was used. The original method consists of estimating item parameters in a single estimation run using the multiple data sets. The linear scale indeterminacy for this combined estimation run can be resolved by either standardizing sample distribution or IRT item parameters. The item parameter estimates for all three sets of items are automatically on the same scale since they were jointly obtained subject to whatever linear constraints were imposed to set the unit for that estimation run. A variant of the original procedure is to use known parameters for a subset of items, in this case the IALS item parameters. By treating the IRT parameters of IALS items as fixed, the jointly estimated item parameters are linked in the most powerful way—at the item level. However, we cannot just assume that the original item parameters fit the new data. It is crucial to examine the invariance of the IALS item parameters for the TOWES field study.

Identical item calibration procedures, described in detail in the IALS technical report (Murray, Kirsch & Jenkins, 1998), were carried out separately for each of the three literacy scales. Using Yamamoto's HYBIL(1989) computer program based on the same EM algorithm, the two-parameter logistic IRT model was fit to each item.

The IALS cognitive items were estimated in 1994 using over 25,000 adults and further validated in 1996 and then again in 1998. The method of parameter calibration in effect puts all survey results onto a designated scale. The English speaking Canadian set of IALS item parameters was used to link TOWES scales.

The IALS item parameters calibrated for the TOWES linking study must fit well in order to justify the use of the item parameter estimates without modification. A graphical method as well as a χ^2 statistic was used to verify such fit. The statistic indicated a very good fit. Only seven out of the 96 IALS items across the three literacy scales received new item parameters due to misfit of the Canadian IALS item parameters. This is less than 10% of the items and is typical of what we have seen in international adult surveys. On average, the root mean squared deviation of observed proportion correct from predicted proportion correct was 0.061 for Prose, 0.054 for Document, and 0.046 for Quantitative literacy scale items. Mean deviation on IALS items were seldom above 0.1 and mean is 0.0. A detailed table containing the deviation statistics for each item by scale is presented in the appendix to this report. Our conclusion is the fit of the IALS item parameters to the TOWES data support the use of the original transformation constants without major modification.

To obtain stable parameter estimates, proficiency distributions for the sample were estimated during calibration, i.e., not fixed. It is known that the samples for each assessment came from somewhat different populations with different characteristics. The calibration procedure should take into account the possibility of systematic interaction of samples and items to estimate unbiased estimates of sample distributions and item parameters. For that reason, multinomial distribution for the population was estimated concurrently with item parameters.

The following table presents the number of items with item parameters common to IALS and the number of items in which new item parameters had to be estimated. Eighty-nine items out of 96 were able to retain the original IALS item parameters, thus assuring the common scale representation by TOWES and IALS items.

Summary of IALS Item Parameters Which Were Retained on Each Literacy Scale

	IALS	Estimated	Total
Prose	32	1	33
Document	30	2	32
Quantitative	27	4	31
Total	89	7	96

Estimated item parameters and classical test statistics of items are presented in the appendix. Since the TOWES items are linked at the item level through calibration, IRT parameters are still on the provisional scale, ie. not ready to be reported on the IALS scales. Because of the success in retaining the original IALS parameters, the transformation constants that put the new TOWES literacy items onto the IALS scales are identical to those used for the IALS and are listed here.

Transformation Constants can be Applied to Provisional Scale to Produce Reported Scale

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

SUMMARY AND CONCLUSION

This report summarizes the procedures that were followed both in the design of the field-test study and the analyses that were undertaken to scale and link the TOWES literacy items to IALS. The range of analyses that were undertaken from estimating scoring reliability to estimating item order effects to scaling and linking all yield consistent evidence to support the reliability and validity of the TOWES items and for placing them on the IALS literacy scales. A summary of the issues that were addressed analytically and the outcomes that were observed are presented here. While there are a few items that should be dropped from future use, the majority of the items fit the model quite well and can be used to assess individuals for the purpose of estimating their literacy proficiencies as was done in IALS.

Summary of Analytic Procedures and Outcomes

Issue Addressed	Need for addressing issue	Outcome observed
Scoring reliability or inter-rater agreement	Minimize the component of measurement error due to inter-rater agreement	High inter-rater agreement was found indicating that this component of the survey contributed little to measurement error
Block order effects	Estimate measurement error coming from position effects of items in blocks	No position effects therefore no new variable had to be taken into account in estimating the model
Treatment of Missing Data	Distinguished omitted from not reached items and treats not reached as unrelated to ability	Reduces bias in the data
Parameter estimation and linking	Linking done at item level to maintain scale dimensionality	Linking and parameter estimation was highly successful resulting in TOWES items being estimated on IALS scales

REFERENCES

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

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46*, 443-449.

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

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society (Series B), 39*, 1-38.

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

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*, 359-381.

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

Murray, S., Kirsch, I. & Jenkins, L. Adult Literacy in OECD Countries: Technical report on the First International Adult Literacy Survey**.** Washington, DC: National Center for Education Statistics, 1998.

Organisation for Economic Co-operation and Development, The Well-being of Nations: The Role of Human and Social Capital, Paris, 2001

Organisation for Economic Co-Operation and Development and Statistics Canada, Literacy in the Information Age: Final Report of the International Adult Literacy Survey, Ottawa, Canada, 2000.

Sum, A; Kirsch, I. & Taggart, R. The Twin Challenges of Mediocrity and Inequality: Literacy in the US from an International Perspective. Princeton, NJ: Educational Testing Service, 2002.

Tuijnman, A. Benchmarking Adult Literacy in America: An International Comparative Study. Jessup, MD: US Department of Education, September, 2000.

Willms, J.D. Inequalities in Literacy Skills Among Youth in Canada and America. Ottawa, Canada: Statistics Canada, September, 1999

Yamamoto, K. & Muraki, E. (1991). Non-linear transformation of IRT scale to account for the effect of non-normal ability distribution on item parameter estimation. A paper presented at the annual 1991 American Educational Research Association meeting, Chicago, IL. 1991.

Yamamoto, K (1989) HYBIL, Computer program to estimate parameters of Hybrid model.

Yamamoto, K. (1998). Scaling and Scale Linking. In T.S. Murray, I.S. Kirsch, & L. Jenkins. *Adult Literacy in OECD Countries.* (Technical report on the First International Adult Literacy Survey). National Center for Education Statistics. Washington, DC: U.S. Department of Education.