Maintaining Comparability: The Move to a Single Literacy Score for the OSSLT

Report Prepared for the Education Quality and Accountability Office (EQAO) by

Nizam Radwan, Ph.D.
Psychometrician, EQAO

Mark D. Reckase, Ph.D.
Scholar in Residence, EQAO; Professor, Michigan State University

W. Todd Rogers, Ph.D.
Scholar in Residence, EQAO; Professor, University of Alberta

APRIL 2010
About the Education Quality and Accountability Office

The Education Quality and Accountability Office (EQAO) is an independent provincial agency funded by the Government of Ontario. EQAO's mandate is to conduct province-wide tests at key points in every student's primary, junior and secondary education and report the results to educators, parents and the public.

EQAO acts as a catalyst for increasing the success of Ontario students by measuring their achievement in reading, writing and mathematics in relation to Ontario Curriculum expectations. The resulting data provide a gauge of quality and accountability in the Ontario education system.

The objective and reliable assessment results are evidence that adds to current knowledge about student learning and serves as an important tool for improvement at all levels: for individual students, schools, boards and the province.

About EQAO Research

EQAO undertakes research for two main purposes:

• to maintain best-of-class practices and to ensure that the agency remains at the forefront of large-scale assessment and
• to promote the use of EQAO data for improved student achievement through the investigation of means to inform policy directions and decisions made by educators, parents and the government.

EQAO research projects delve into the factors that influence student achievement and education quality, and examine the statistical and psychometric processes that result in high-quality assessment data.
Abstract

The Ontario Secondary School Literacy Test (OSSLT) was changed in March 2006, when it was shortened in response to recommendations in an external review of the EQAO assessment program. The decision was also made at that time to report the results as a single literacy score rather than as separate reading and writing scores. Because of these changes, the OSSLT administered in March 2006 was not parallel to that administered in October 2004. However, the desire remained to link the 2004 and the 2006 OSSLTs so as to yield results that enable the measurement of change in student performance between 2004 and 2006. Two 2004 “test forms” were considered as means for establishing the link: the actual, full 2004 test, scored to yield a single literacy score, and a pseudo-test, constructed from items included in the 2004 test and designed to mimic the 2006 test. Linking the 2006 OSSLT with these two test forms resulted in similar percentages of successful students. The pseudo-test form was selected because it was a better match with the 2006 OSSLT, and the linked scores were treated as comparable rather than equated.
Introduction

The Ontario Secondary School Literacy Test (OSSLT) measures the reading and writing skills of Grade 10 students attending public and private schools in the province of Ontario. Passing the OSSLT is one of the provincial high school graduation requirements. Students who write the OSSLT once and are unsuccessful are entitled to fulfill the literacy requirement by passing the Ontario Secondary School Literacy Course (OSSLC).

The OSSLT was first administered in 2002. Two half-days were required to complete it until 2006. Both reading and writing items were assessed on both days. The OSSLT used a conjunctive decision model: a score for reading and a score for writing were reported, and students were required to pass both reading and writing in order to fulfill the literacy requirement.

The OSSLT was shortened in 2006 in response to recommendations from a comprehensive external review of all EQAO assessment programs (Wolfe, Childs, & Elgie, 2004). The new test, which is approximately half the length of the original, requires one half-day. Furthermore, the conjunctive decision model has been replaced by a compensatory decision model. One literacy score is reported rather than separate scores for reading and writing. The reading and writing components are approximately equally weighted so that each counts for approximately half the total score. This was done to mirror the previous decision model, which explicitly weighted reading and writing equally.

With the changes between 2004 and 2006, the desire remained to link the 2006 OSSLT with the 2004 OSSLT so that the change in student performance between 2004
and 2006 could be assessed. There are, however, several prerequisite conditions that need to be satisfied before two test forms can be considered equated (Kolen & Brennan, 2004).

First, the forms should be construct-equivalent. That is, it should be possible to interpret the scores yielded by the forms validly in terms of the same construct or domain. Additionally, the means, standard deviations and reliabilities should be the same and the correlation corrected for attenuation due to unreliability should be 1.00 between the two forms.

Second, according to the equity condition, for each group of examinees of identical ability or with equal achievement, the conditional distribution of scores on form X (after equating) should be the same as the conditional distribution of scores on form Y. For each group of students with the same true score, \( \tau \), the conditional frequency distribution of scores on form X (after equating) should be the same as the conditional frequency distribution of scores on form Y:

\[
E_{j/\tau}[eq_Y(X): \tau] = E_{j/\tau}(Y/\tau),
\]

where \( E_{j/\tau} \) is the expectation taken over \( j = 1, 2, \ldots, N_\tau \), \( N_\tau \rightarrow \infty \) students with true score \( \tau \),

\( eq_Y \) is an equating function that transforms the scores on form X to the score scale of form Y,

\( (X): \tau \) is the distribution of transformed X scores for the students with true score \( \tau \) and

---

1 The OSSLT for the 2004–2005 school year was administered in October 2004, while that for 2005–2006 was administered in March 2006. No test was administered in 2005.
$(Y / \tau)$ is the distribution of Y scores for the students with true score $\tau$.

Third, there should be population invariance. The equating transformation for taking the scores on X onto the Y score scale should be the same for all samples from the same population.

Fourth, the symmetry condition requires that the transformation from X to Y be the same as that for Y to X. This differs from regression, where the regression of Y on X is not necessarily the same as the regression of X on Y.\(^2\)

The first of the four conditions, construct equivalence, is of particular interest in the present study given the change in OSSLT length and the adoption of a compensatory decision-making model. Therefore, the first purpose of this study was to determine whether the 2004 and 2006 OSSLTs were construct-equivalent. If they were, then the scores could be equated. If they were not, then depending on the degree of similarity, the two score distributions could be scaled for comparability (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

The second purpose of this study was to assess the linking process for equating or scaling. As mentioned above, in 2004, the hypothetical construct of literacy was estimated using two subtests, reading and writing. While students were assumed to be ordered along a single literacy continuum, the tests produced two separate estimates of literacy and both were used to estimate students’ locations by determining whether their scores were above the respective cut scores on the two tests. In 2006, a single score was

\(^2\) If the means and standard deviations of the two forms are respectively equal, then the regression of Y on X and X on Y will be the same.
used for the pass/fail decision. That score, which is based on a combined scoring of reading and writing test items, should be a closer approximation to the hypothetical construct of literacy than that of the 2004 OSSLT. Given this assumption is met, a complication in the linking process arises since the 2004 OSSLT does not have the same weighting of reading and writing as the 2006 OSSLT. In this case, the difference in weightings must be taken into account during the linking process. Therefore, an additional purpose of this study was to investigate how to account for such difference during the linking process when the 2006 test was considered to be the operational definition of the construct.

**Examination of Construct Equivalence of the 2004 and 2006 OSSLTs**

**Definition of Literacy**

The definition of literacy adopted for the OSSLT is as follows:

For the purpose of the OSSLT, literacy comprises the reading and writing skills required to understand reading selections and communicate through a variety of written forms as expected in *The Ontario Curriculum* across all subjects up to the end of Grade 9. (Education Quality and Accountability Office, 2007, p.10).

This definition has not changed since the OSSLT was introduced in 2002.

This definition is expanded by describing what the students are expected to use and do when responding to the OSSLT. For reading, the students must use strategies to interact with a variety of narrative, informational and graphic text selections to construct and gain an understanding of the meaning of texts of different forms, demonstrate their understanding of explicit and implicit meaning and connect their understanding of what
they have read to their own personal experience and knowledge. For writing, students are prompted to write two short responses, a series of paragraphs expressing and supporting an opinion they formulate from the prompt and a news report they develop from the prompt. Through their responses, students demonstrate their ability to communicate ideas and information clearly and coherently (Education Quality and Accountability Office, 2007, p.10).

*Comparability of the 2004 and 2006 OSSLTs*

Both the 2004 and the 2006 OSSLTs were constructed to assess the skills outlined in the definition of literacy and its expansion. However, the 2006 OSSLT was approximately half the length of the 2004 OSSLT and had a single literacy score, rather than separate reading and writing scores. According to the selection criteria, the 2004 and 2006 items (the reading selections and accompanying multiple-choice and open-response items and the writing prompts for the English- and French-language forms) needed to be similar in the following respects:

a. levels of cognitive processing and employment of reading and writing strategies;

b. difficulty levels of the reading passages and

c. reading and writing item difficulty.

(Although the OSSLT and the *Test Provincial de Compétences Linguistiques* (TPCL), the French-language literacy test, are constructed separately, the number of items and points are the same.)

The first two rows of the two panels in Table 1 contain, respectively, a description of the 2004 and 2006 OSSLTs in terms of the number of items and points for each of the
three reading skills and for writing, which together constitute literacy. While the reading selections and extended-response writing prompts used in 2004 and 2006 measured the same general skills, the tests differed somewhat in their distribution of multiple-choice reading items, short-answer reading items and short-answer writing items. The 2006 operational form contained eight multiple-choice items that measured writing conventions and grammar, while the 2004 form did not.

Furthermore, while the multiple-choice items in the 2004 and 2006 tests were construct-equivalent, the construct equivalence of the prompts for the short reading items was more tenuous. For example, for one pair of items, the students in 2004 were to provide an explanation while the students in 2006 were to provide a summary. The two short-answer writing items used in 2006 were paired with a longer writing item in 2004 that required students to summarize a long reading passage.

In totality, the degree of fit between the 2004 and 2006 tests was not strong enough to consider them construct-equivalent. The differences in item distribution and the lack of fit between open-response items precluded such a claim. Therefore, scaling for comparability was adopted (American Educational Research Association, et al., 1999).
### Table 1

**Distribution of Items and Points Across Reading and Writing by Year**

<table>
<thead>
<tr>
<th>Year</th>
<th>Reading Skill</th>
<th>Writing Skill</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of Items</td>
<td>Writing Skill</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R1</td>
<td>R2</td>
<td>R3</td>
</tr>
<tr>
<td>2004</td>
<td>200</td>
<td>21</td>
<td>24</td>
</tr>
<tr>
<td>%</td>
<td>14.0</td>
<td>15.1</td>
<td>20.2</td>
</tr>
<tr>
<td>2006</td>
<td>220</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>%</td>
<td>29.3</td>
<td>0.0</td>
<td>29.3</td>
</tr>
<tr>
<td>2004 PT</td>
<td>220</td>
<td>14</td>
<td>21</td>
</tr>
<tr>
<td>%</td>
<td>30.6</td>
<td>0.0</td>
<td>42.9</td>
</tr>
<tr>
<td></td>
<td>Number of Points</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>200</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>%</td>
<td>9.9</td>
<td>14.9</td>
<td>14.9</td>
</tr>
<tr>
<td>2006</td>
<td>200</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>%</td>
<td>14.6</td>
<td>0.0</td>
<td>14.7</td>
</tr>
<tr>
<td>2004 PT</td>
<td>200</td>
<td>14</td>
<td>21</td>
</tr>
<tr>
<td>%</td>
<td>14.0</td>
<td>0.0</td>
<td>21.0</td>
</tr>
</tbody>
</table>

**Notes:**
- *k* and *n* are the number of items and points, respectively; % is the corresponding percentage.
- **R1** – Understanding explicitly stated ideas and information.
- **R2** – Understanding implicitly stated ideas and information.
- **R3** – Making connections between ideas and information in a reading selection and personal knowledge and experience.
- **MC** – Multiple-choice items.
- **OR** – Open-response reading items.
- **SW** – Short-writing items.
- **LW** – Long-writing items.
- **PT** – Pseudo-test.

The difference between the number of items and the number of points for R1, R2 and R3 is due to the use of multiple-choice items, which were dichotomously scored (0, 1), and open-response items, which were polytomously scored (0, 1, 2).

Points are awarded for both topic development and writing conventions.
Scaling for Comparability

Although the hypothetical construct is usually thought of as existing before a test is constructed, the construct is often operationally defined by the first form of a test to be produced. When a test measures achievement, the construct is very complex and is typically a weighted combination of specific skills and knowledge. For example, literacy is a weighted combination of vocabulary knowledge, knowledge of sentence structure, reading comprehension, background knowledge, knowledge of grammar and usage, organizational skills, spelling knowledge and many other components. The weighting of these different components can be done explicitly (for example, reading and writing can be stated to have equal weights) or it can be determined by the number of items of each type included in the test and the quality of information provided by the items. How item responses are combined to form a score has an influence on the weighting of components.

If a specific form of a test can be assumed to be well designed and constructed, then the reported score on the test is expected to reflect the desired weighting of components and to be highly related to the construct. That means that the ordering of students using the reported score should be strongly related to the ordering that would be observed if the students’ locations on the construct itself were available. The implication of this observation is that the goal of scaling for comparability is to estimate the location of the pass/fail point on the test form reporting scale (e.g., in 2006) that corresponds to the location on the construct implied by previous decisions (e.g., in 2004).

A complication in the process of scaling for comparability is that earlier versions of a test may not have the same weighting of component parts as the current version that is being used as the definition of the construct. That means the differences in weightings
will have to be taken into account when scaling for comparability. In the case considered here, the 2006 operational form was assumed to reflect the weighting of the reading and writing components for the EQAO literacy construct accurately. The 2004 form had two separately scored parts, so the weighting of components was handled differently. Consequently it was necessary to consider this difference when scaling the 2004 and 2006 forms for comparability.

The Psychometric Expert Panel for EQAO provided recommendations for the data analysis and psychometric procedures for the transition from 2004 to 2006 that would enable linking the 2004 and 2006 OSSLT scores. The panel recommended two test models that reflected different weightings of the reading and writing components to scale the 2004 and 2006 OSSLTs. The first was the total test model, in which the 2004 test with its two separate components was treated as one test yielding one score. The second was the pseudo-test model, which involved creating a test form “parallel” to the 2006 OSSLT using the items included in the 2004 OSSLT. The purpose of the second stage of this study was to compare these two test models to determine which model yielded the superior scaling results.

Test Models for Scaling

Total Test Model

The total test model was based on the assumption that the full set of items administered operationally in 2004 provided the appropriate weighting of the reading and writing skills for the literacy construct, as defined, when the two components were analyzed together as
a single test. That is, the weighting provided by the number of score points in the full set of items matched the weighting implied by the 2006 form of the test when the reading and writing items were calibrated together.

**Pseudo-Test Model**

In 2006, a single score was used for the pass/fail decision. That score was based on the combined scoring of the reading and writing items. Thus, the single score should be a closer approximation of the literacy construct than the 2004 test with its separate reading and writing parts.

Because of these changes, a 2004 pseudo-test was constructed from the 2004 operational items that best “mimicked” the 2006 operational test. The type and number of items selected was to match as closely as possible the type and number of items and the total points for each of the three reading skills and for writing in the 2006 OSSLT. Attempts were made to ensure that the items selected from the 2004 operational form satisfied these criteria better than the entire 2004 form. The three long-writing items included in the pseudo-test were scored with a four-point rubric in 2004 while the two long items included in the 2006 operational form were scored with a six-point rubric. Consequently, to obtain a score weight equal to the score weight of the 2006 writing items, the three 2004 long-writing items’ scores were multiplied by four.

The number of items and score points for the 2004 pseudo-test are provided in the last row of each panel of Table 1. Comparing the values in these two rows with the corresponding rows for the 2004 and 2006 OSSLTs reveals that the pseudo-test provided

---

3 Field-test items are embedded in each OSSLT following a matrix sampling plan. Consequently there is more than one form in each year. Each form contains the same operational items and a set of field-test items. The items for the next year’s form are selected from the embedded field-test items and these are the items used to link the two years.
a closer fit to the 2006 OSSLT than the total test. Therefore, construction of a 2004 pseudo-test from the operational items contained in the 2004 OSSLT to mimic the 2006 OSSLT was deemed successful.

**Calibration and Scaling Sample**

A set of exclusion rules was implemented for the selection of the calibration and scaling sample to ensure that the sample was representative of the provincial population. First, previously eligible students were removed. Then the following categories of first-time eligible students were excluded from the 2004 and 2006 scaling samples:

1. students with no work or incomplete work in a major section of the test;
2. students receiving accommodations, except in the case of extra time;
3. students who were exempted, deferred or taking the Ontario Secondary School Literacy Course (OSSLC) and
4. students who were home schooled.

After the exclusion criteria were applied, the numbers of first-time eligible students in the OSSLT scaling samples were 137,496 and 4645 for the 2004 OSSLT and TPCL, respectively, and 146,280 and 5009 for the 2006 tests. The number of French students was considerably smaller than the number of English students.

**Calibration**

A modified Rasch model (Rasch, 1960, 1980) was used to calibrate the multiple-choice items, and the modified partial credit model (PCM) (Masters, 1982) was used to calibrate the open-response items. These models were selected for the OSSLT/TPCL because of the small number of French-language students. Due to the number of versions of the tests with different field-test items embedded in the common set of operational items, the
number of French students who responded to each field-test item varied from about 400 to 600. In contrast, the number of English students who responded to each field-test item was approximately 7000, with approximately 1500 student responses to open-response items scored for scaling. However, EQAO’s policy was to keep the analysis procedure the same for French- and English-language students. Consequently, due to insufficient sample size for the French field-test items, more complex models such as the two- and three-parameter models were not used for the tests.

The calibrations were completed using PARSCALE 4.1 (Muraki & Bock, 2003). The \( a \)-parameter for the multiple-choice and open-response items was set to 0.588. The pseudo-guessing parameter for the multiple-choice items was set to 0.20 \( \left( \frac{1}{k + 1} \right) \); where \( k \) is the number of options) to account for the possibility of students with a very low ability correctly answering an item. Thus, the modified Rasch model is given by

\[
P_i(\theta) = 0.20 + (1 - 0.20) \frac{\exp^{D \cdot 0.588(\theta - b_i)}}{1 + \exp^{D \cdot 0.588(\theta - b_i)}},
\]

where \( P_i(\theta) \) is the probability of a student (with ability \( \theta \)) answering item \( i \) correctly;

\( D \) is a scaling factor = 1.7;

and

\( b_i \) is the difficulty parameter for item \( i \).

The PCM was used to estimate the item and ability parameters for the open-response items. The PCM is given by
where \( P_{ix}(\theta) \) is the probability of a student (with ability \( \theta \)) receiving a particular score \( x \) on item \( i \),

and

\( \delta_{ij} \) is the item \( i \) step difficulty parameter associated with a category score \( j \).

The first analysis performed was to check the psychometric quality of all 2004 field-test items prior to their inclusion in the calibration of the 2006 operational items and the linking of 2006 and 2004 test scores. This check was completed in three steps:

1. Calibrate the 2004 operational items and the field-test items selected for the 2006 operational test.
2. Calibrate the 2006 operational items.
3. Compare the item parameters of items common to 2004 and 2006. Plots of the pairs of item parameters should appear along a straight diagonal line. Items corresponding to points outside the 95% confidence band around this line are considered to be outliers (Hu, Rogers, & Vukmirovic, 2008) and are dropped from scaling.

Based on this analysis, all field-test items were retained.

**Scaling Procedure**

The forward-fixed parameter common-item non-equivalent group design was used to scale the 2004 and 2006 OSSLT scores. Common items were selected from the 2004 field-test items and used as 2006 operational items. To place the 2006 and 2004 scores on
the same scale, the item parameters from the 2006 calibration were fixed when the parameters of the 2004 assessment were recalibrated.

The process used to place the 2004 and 2006 OSSLTs on a common scale is diagrammed in Figure 1 for the total test model and Figure 2 for the pseudo-test model. The steps associated with each procedure are outlined below the appropriate diagram.

Figure 1
*Calibration Procedure for the Total Test Model*

**Scaling with the Total Test**

1. Calibrate all of the 2006 operational items using all of the operational data.
   a. Create a data file that contains the data for the 2006 items that were scored operationally for the scaling sample.
   b. Run the IRT calibration program and save both the item parameters for each item and the estimates of proficiency for each examinee. Use the
item parameters to scale the 2004 scores onto the 2006 scale. Use the
distribution of proficiency to determine the proportion of students who
passed in 2006.

2. Identify the items from the 2006 operational item set that were exactly the same in
the 2004 field test. Use only identical items to link the 2004 results to the 2006
scale.

3. Recalibrate the 2004 reading and writing tests as a single literacy test, and place
the scores on the 2006 scale.
   a. Create a data file that contains
      i. the examinee responses to the 2004 operational items;
      ii. the examinee responses to the 2004 field-test items that were
          brought forward unchanged to the 2006 form and
      iii. a pass/fail decision for each examinee.
   b. Apply exclusion rules to the 2004 data file to obtain the data for
      calibration.
   c. Run the IRT calibration program, fixing the item parameters for the field-
      test items brought forward to the 2006 operational forms at the values
      from the calibration in step 1b. All of the operational items from the 2004
      calibration now have item parameters on the 2006 scale.

4. Use the item parameter estimates for the 2004 operational items from step 3c to
compute Rasch proficiency estimates for the full set of 2004 operational data.
   a. Fix the item parameters at the values estimated in step 3c.
   b. Estimate the proficiency level for each student in the calibration sample.
5. Determine the Rasch proficiency value so that the proportion of students above it equals the proportion of students who passed in 2004.

   a. Tally the pass rate for the examinees identified in step 2.

   b. Determine the point in the distribution of proficiency estimates in step 4b with the same proportion above it as the pass rate in step 5a.

   c. Use the value from step 5b as the pass/fail cut score in the 2006 proficiency distribution (see step 1b).

The full set of operational items from 2004 was used to define the Rasch proficiency scale. Consequently, it is likely that the pass/fail score in step 5c was highly reliable.

Figure 2

*Calibration Procedure for the Pseudo-Test Model*
**Scaling with the Pseudo-Test**

1. Calibrate all of the 2006 operational items using all of the operational data.
   a. Create a data file that contains the data from the items that were scored operationally for the 2006 scaling sample.
   b. Run the IRT calibration program and save both the item parameters for each item and the estimates of proficiency for each individual. Use the item parameters to place the 2004 results on the 2006 scale. Use the distribution of proficiency to determine the proportion of students who passed in 2006.

2. Identify the items from the 2006 operational item set that were exactly the same in the 2004 pseudo-test.

3. Calibrate the 2004 pseudo-test as a single test on the 2006 scale.
   a. Create a data file that contains
      i. the results from 2004 pseudo-test items and
      ii. the 2004 field-test items that were brought forward unchanged to the 2006 form.
   b. Apply exclusion rules to the 2004 pseudo-test data file to obtain the data for calibration.
   c. Run the IRT calibration program, fixing the item parameters for the field-test items brought forward to the 2006 operational forms at the values from the calibration in step 1b above. All of the pseudo-test items from the calibration now had item parameters on the 2006 scale.
4. Use the item parameter estimates for the pseudo-test items from step 3c to compute Rasch proficiency estimates for the full set of operational data in 2004.
   a. Fix the item parameters at the values estimated in step 3c.
   b. Estimate the proficiency level for each student in the sample.
5. Determine the Rasch proficiency value with the same proportion above it as the 2004 pass rate.
   a. Tally the pass rate for the examinees used in step 4b.
   b. Determine the point in the distribution of proficiency estimates in 4b that has the same proportion above it as the pass rate in step 5a.
   c. Use the value from step 5b as the pass/fail cut score in the 2006 proficiency distribution.

**Results and Discussion**

The scaling results based on the linking samples are summarized in Table 2 (English-language students appear in the top panel and French-language students in the bottom panel). The percentages of successful English- and French-language students in the 2004 scaling samples were 88.6% and 86.3% respectively. The percentage of English-language students successful in the 2006 scaling sample was 89.2% for the total test and 87.8% for the pseudo-test. The percentage of French-language students successful in the 2006 scaling sample was 85.8% for the total test and 86.4% for the pseudo-test.

The total test model resulted in an increase of 0.6 percentage points in the 2006 English-language scaling sample and a decrease of 0.5 percentage points in the 2006 French-language scaling sample compared to the percentage of successful students in the
2004 scaling sample. The pseudo-test model resulted in a decrease of 0.8 percentage points in the 2006 English-language scaling sample and an increase of 0.1 percentage points in the 2006 French-language scaling sample. Despite the difference in sample size, the English and French scaling results are comparable.

Table 2

*Scaling for Comparability Results for the Total Test and Pseudo Test: English and French Language*

<table>
<thead>
<tr>
<th></th>
<th>Percentage of Successful Students</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test model for scaling</strong></td>
<td><strong>2004</strong></td>
</tr>
<tr>
<td><strong>English-language students</strong></td>
<td></td>
</tr>
<tr>
<td>Total test</td>
<td>88.6</td>
</tr>
<tr>
<td>Pseudo-test</td>
<td>88.6</td>
</tr>
<tr>
<td><strong>French-language students</strong></td>
<td></td>
</tr>
<tr>
<td>Total test</td>
<td>86.3</td>
</tr>
<tr>
<td>Pseudo-test</td>
<td>86.3</td>
</tr>
</tbody>
</table>

The total test and the pseudo-test models yielded similar results for the 2006 OSSLT scaling samples. However, each has relative advantages and disadvantages in its implementation. In the case of the total test, while avoiding the need to take the time to construct a pseudo-test from a fixed set of previously administered items is an advantage, the fact that the weighting of the skills and knowledge in the Year 1 test (implied by the score points) may not match that for the Year 2 test is a possible disadvantage. Inconsistent weighting between the two years may result in an unintended shift in the proportion passing in Year 2 (see Schaeffer, Henderson-Montero, Julian, & Bene, 2002; Sykes & Hou, 2003; Rogers & Nowicki, 2009).
The pseudo-test should better correspond to the construct measured in the second year. However, the pseudo-test needs to be constructed, and there may not be a sufficient number of items to match the current specifications for the construct to be measured. Further, the pseudo-test will contain fewer items than the full set, which may reduce the reliability of the test scores and adversely affect the scaling results. In the present case, the values of Cronbach’s alpha were 0.82 and 0.81, respectively, for the English- and French-language pseudo-tests and 0.86 for both the English- and French-language operational tests.

**Conclusion**

The pseudo-test model was chosen to scale the 2004 and 2006 OSSLTs for comparability, after the number of items was reduced by approximately one-half and the conjunctive decision-making model was replaced with a compensatory decision-making model. The pseudo-test was more similar to the 2006 operational test specifications and proportion of items by type and skill than was the total test. While it appeared that the 2004 total test model provided an appropriate weighting of reading and writing components relative to the literacy construct, the pseudo-test model explicitly provided more appropriate weighting.

One of the major challenges in large-scale assessments is accommodating changes in the test that need to be made in response to changes in curriculum, administration time or decision-making model. Despite these changes, the desire remains to know whether the performance of students changed from year to year. The pseudo-test model outlined in this paper provides a viable method of scaling for comparability when there is a change in the test structure, length of administration or reporting outcomes.
References

American Educational Research Association, American Psychological Association, &
Research Association.

Literacy Test*. Toronto, ON: Author.

methods in the presence of outlier common items. *Applied Psychological

Kolen, M. J., & Brennan, R. L. (2004). Test equating, scaling, and linking: Methods and


174.


rating-scale data. (Version 4.1)* [Computer Software] Chicago, IL: Scientific
Software International.

(Copenhagen, Danish Institute for Educational Research), expanded edition


