EQAO

Education
Quality and
Accountability
Office

The October 2002 Ontario Secondary School Literacy Test: Technical Paper on the Results of the Reliability and Validity Processes

The October 2002 Ontario Secondary School Literacy Test: Technical Paper on the Results of the Reliability and Validity Processes

This technical paper is a companion piece to the Education Quality and Accountability Office's (EQAO's) October 2002 Ontario Secondary School Literacy Test: Summary Paper on Quality Assurance Measures and Reliability and Validity Processes. In addition to the information provided in the summary paper, this paper provides detailed information on the reliability and validity of the October 2002 Ontario Secondary School Literacy Test (OSSLT). Unlike the reader of the summary paper, the reader of this paper is assumed to have some familiarity with advanced topics in measurement and statistics.

A. Quality Assurance Measures for Marker Consistency

Before the OSSLT was scored, all the student papers received from the schools were scrambled. This ensured that the papers were distributed randomly to the markers, so that no one marker scored a disproportionate number of student papers from any one school. In addition, *all* the student work used in the training of markers and during the actual scoring underwent blind scoring. *Blind scoring* means that there is no information on the student work that will allow a marker to identify a student, a school or a school board.

The key to the accurate scoring of students' papers is the appropriate training of the markers. Throughout the marking of the October 2002 OSSLT, EQAO took a number of steps to ensure that all markers were well trained and monitored and that they scored student papers in a similar way. These steps included the use of training papers, a marker readiness test, a chain-marking method, group-marking papers and double marking. Each of these steps is described in the summary paper, available at www.eqao.com.

B. Reliability

During the marking of the October 2002 OSSLT, a random selection of student papers was chosen to be reinserted into the marking flow to receive a second set of marks. These reinserted papers, approximately 900 English-language student papers and 900 French-language student papers, represented a wide spectrum of student performances. Following the marking, the data derived from the original marking and the reinsertion were compared to obtain an index of inter-rater reliability (marker agreement).

The total scores for each student from the original marking and the reinsertion were correlated to derive an inter-rater reliability index. The inter-rater reliability results show the overall inter-rater reliability coefficients to be over 0.90 for both components of the test and for both languages. Specifically, in the reading component, the inter-rater reliability coefficient was 0.99 (n = 882) for English-language papers and 0.93 (n = 875) for French-language papers. For the writing component, the inter-rater reliability coefficient was 0.94 (n = 900) for English-language and 0.93 (n = 898) for French-language papers. These results indicate that student responses were scored consistently, that the scores are reliable and that markers applied the scoring rubrics in a uniform manner.

For large-scale assessments, inter-rater reliability coefficients are typically in the mid to high 0.80s or 0.90s (refer to the *Arizona's Instrument to Measure Standards' [AIMS]' Guide to Understanding AIMS Results;* the Massachusetts Comprehensive Assessment System: 1998 Technical Report [1999] and the Washington Assessment of Student Learning: Grade 10, 1999 Technical Report, 2001). Taken together, the results from the October 2002 OSSLT are in keeping with the standards concerning the inter-rater reliability of large-scale assessments.

C. Validity

The Content Representation of the Test

An important aspect of the validity of the OSSLT is the fact that the test items and tasks are derived from the larger curriculum content domain on which the test is based. All reading selections and writing tasks on the OSSLT are based on the knowledge and skills covered in *The Ontario Curriculum* across all subjects up to the end of Grade 9. Consequently, students are expected to demonstrate cumulative content knowledge and skills across all subjects in *The Ontario Curriculum.* The recently published Curriculum Connections (2003) show some of the correspondence between the OSSLT content and *The Ontario Curriculum*.

Assessing the Internal Validity of the Test

To assess the internal validity of the OSSLT, the following aspects are examined: a) the relationships within each component of the test, b) the relationship between the two components of the test and c) the consistency of student responses across the different parts of the test.

The following analyses are based on data pertaining to newly eligible students, that is, students taking the test for the first time and who were present on both days of the test.

a) Relationships Within Each Component of the Test

Reading Component

To assess the internal validity of the reading component, the relationships within and among the different selection types, reading skills and question types are examined. This component is composed of 12 reading selections, divided into three selection types: i) information, ii) graphic and iii) narrative. For each selection type, students are assessed on three reading skills: i) understands directly stated ideas and information, ii) understands indirectly stated ideas and information and iii) makes connections between personal experiences and the ideas and information in the reading selections. There are 40 multiple-choice questions, 35 short-answer questions and 25 questions requiring an explanation.

For the English-language test, the correlations among the three selection types range between 0.73 and 0.80; the correlations among the three reading skills are between 0.75 and 0.81; and the correlations among the three item types are between 0.72 and 0.79.

For the French-language test, the correlations among the three selection types range between 0.72 and 0.78; the correlations among the three reading skills are between 0.76 and 0.83; and the correlations among the three item types are between 0.74 and 0.83.

Writing Component

For the writing component, the analysis involves examining the relationships among the four writing tasks and between each of these tasks and the overall performance. The four tasks are a summary, a series of paragraphs expressing an opinion, a news report and an information paragraph.

For the English-language test, the correlations between each of the four tasks and the overall student performance range between 0.65 and 0.69. These correlations show that students with high performances across the four tasks are likely to succeed on the writing component of the test. The correlations among the four writing tasks are relatively moderate, ranging from 0.21 to 0.40.

For the French-language test, the correlations between each of the four tasks and the overall student performance range between 0.70 and 0.73. These correlations show that students with high performances across the four tasks are likely to succeed on the writing component of the test. The correlations among the four writing tasks are relatively moderate, ranging from 0.29 to 0.41.

For both the English-language and the French-language results, the relatively moderate correlations among the four writing tasks reflect the fact that each writing task requires specific writing skills that may not be shared among the four tasks. However, students are expected to acquire these skills across all subjects up to the end of Grade 9.

b) Relationship Between the Reading Component and the Writing Component

The overall relationship between the reading component and the writing component was moderately strong, with correlations of 0.67 for the English-language test and 0.69 for the French-language test. Generally, students with higher performances in the reading component also demonstrated higher performances in the writing component. These results are similar to the correlations reported between reading and writing scores (0.65) for the Grade 10 Washington assessment program (Washington Assessment of Student Learning: Grade 10, 1999 Technical Report, 2001).

c) Consistency of Student Responses Across the Different Parts of Each Component of the Test

Reading component

To examine the consistency of student performances on the three selection types, the mean performance and standard deviations (SDs) are presented for students who passed and students who did not pass the component.

Text Types

i) For the English-language test, the mean performances on the different selection types of students who **passed** the reading component (*n* = 103 256) were as follows:

• Narrative	45 (SD = 6.14, maximum 60 points)
• Information	64 (SD = 8.32, maximum 90 points)
• Graphic	38 (SD = 5.18, maximum 50 points)

For the English-language test, the mean performances on the different selection types of students who **did not pass** the reading component (*n* = 32 733) were as follows:

• Narrative	28 (SD = 7.67, maximum 60 points)
• Information	42 (SD = 9.44, maximum 90 points)
• Graphic	25 (SD = 6.46, maximum 50 points)

ii) For the French-language test, the means on the different selection types of students who **passed** the reading component (*n* = 4112) were as follows:

• Narrative	43 (SD = 7.07, maximum 60 points)
• Information	70 (SD = 9.44, maximum 90 points)
• Graphic	41 (SD = 5.54, maximum 50 points)

For the French-language test, the means on the different selection types of students who **did not pass** the reading component (*n* = 853) were as follows:

• Narrative	26 (SD = 7.02, maximum 60 points)
• Information	44 (SD = 9.48, maximum 90 points)
• Graphic	26 (SD = 6.51, maximum 50 points)

Skills

i) For the English-language test, the mean performances in the different skill types of students who **passed** the reading component (*n* = 103 256) were as follows:

- Understands directly stated ideas and information
  47 (SD = 5.42, maximum 60 points)
- Understands indirectly stated ideas and information
  68 (SD = 8.20, maximum 90 points)
- Makes connections between personal experiences and the ideas and information in the reading selections
  32 (SD = 5.67, maximum 50 points)

For the English-language test, the mean performances in the different skill types of students who **did not pass** the reading component (*n* = 32 733) were as follows:

- Understands directly stated ideas and information
  33 (SD = 6.95, maximum 60 points)
- Understands indirectly stated ideas and information
  43 (SD = 10.23, maximum 90 points)
- Makes connections between personal experiences and the ideas and information in the reading selections
  19 (SD = 5.82, maximum 50 points)

ii) For the French-language test, the mean performances in the different skill types of students who **passed** the reading component (*n* = 4112) were as follows:

- Understands directly stated ideas and information
  50 (SD = 6.12, maximum 60 points)
- Understands indirectly stated ideas and information
  66 (SD = 9.53, maximum 90 points)
- Makes connections between personal experiences and the ideas and information in the reading selections
  37 (SD = 5.98, maximum 50 points)

For the French-language test, the mean performances in the different skill types of students who **did not pass** the reading component (*n* = 853) were as follows:

- Understands directly stated ideas and information
  33 (SD = 6.70, maximum 60 points)
- Understands indirectly stated ideas and information
  40 (SD = 8.90, maximum 90 points)
- Makes connections between personal experiences and the ideas and information in the reading selections
  23 (SD = 6.16, maximum 50 points)

Question Types

i) For the English-language text, the mean performances on the different question types of students who **passed** the reading component (*n* = 103 256) were as follows:

• multiple-choice questions	62 (SD = 7.23, maximum 80 points)
• short-answer questions	54 (SD = 6.68, maximum 70 points)
• questions requiring an explanation	30 (SD = 5.82, maximum 50 points)

For the English-language text, the mean performances on the different question types of students who **did not pass** the reading component (*n* = 32 733) were as follows:

• multiple-choice questions	43 (SD = 8.43, maximum 80 points)
• short-answer questions	34 (SD = 9.32, maximum 70 points)
• questions requiring an explanation	18 (SD = 5.74, maximum 50 points)

ii) For the French-language test, the mean performances on the different question types of students who **passed** the reading component (*n* = 4112) were as follows:

• multiple-choice questions	62 (SD = 7.92, maximum 80 points)
• short-answer questions	54 (SD = 8.04, maximum 70 points)
• questions requiring an explanation	38 (SD = 5.82, maximum 50 points)

For the French-language test, the mean performances on the different question types of students who **did not pass** the reading component (*n* = 853) were as follows:

• multiple-choice questions	42 (SD = 7.95, maximum 80 points)
• short-answer questions	31 (SD = 8.19, maximum 70 points)
• questions requiring an explanation	22 (SD = 6.39, maximum 50 points)

For both the English-language and the French-language tests, the results of the above analyses on the selection types, skill types and question types, comparing the scores of students who passed the reading component with the scores of those who did not pass the reading component, showed that the former group consistently scored higher on the different parts of the reading test than the latter group.

Writing Component

To examine the consistency of student performances on the four writing tasks, the results are presented for students who passed and students who did not pass the component. The range of possible scores is 0 to 45 points. The results are presented in terms of the mode (the most frequently obtained score). The percentage of students achieving this score is presented in parentheses.

i) For English-language test, the modes of the scores on the different tasks of students who **passed** the writing component (*n* = 115 329) were as follows:

• Summary	35 (45%)
• Series of paragraphs expression an opinion	35 (60%)
• News report	35 (59%)
• Information paragraph	35 (59%)

For the English-language test, the modes of the scores on the different tasks of students who **did not pass** the writing component (*n* = 20 660) were as follows:

• Summary	0 (49%)
• Series of paragraphs expression an opinion	20 (37%)
• News report	35 (43%)
• Information paragraph	0 (50%)

ii) For the French-language test, the modes of the scores on the different tasks of students who **passed** the writing component (*n* = 4394) was as follows:

• Summary	35 (54%)
• Series of paragraphs expression an opinion	35 (49%)
• News report	35 (51%)
• Information paragraph	35 (56%)

For the French-language test, the modes of the scores on the different tasks of students who **did not pass** the writing component (*n* = 571) was as follows:

• Summary	0 (49%)
• Series of paragraphs expression an opinion	20 (37%)
• News report	20 (32%)
• Information paragraph	0 (36%)

For both the English-language and the French-language students who did not pass the writing component, the summary task and information paragraph tasks appear to have been the most challenging, compared with the series of paragraphs expression an opinion and the news report tasks.

For both the English-language and the French-language tests, the results of comparing the modes of the scores of students who passed the writing component with the modes of the scores of students who did not pass the writing component showed that the former group scored consistently higher on the four tasks of the writing component than the latter group.

For the English-language test, scores on the news report yielded the same mode for the students who passed the writing component and the students who did not pass the writing component. However, the percentage associated with the students who passed the writing component (59%) was higher than for the students who did not pass the writing component (43%).

References

Arizona Department of Education. *Arizona's Instrument to Measure Standards (AIMS) Guide to Understanding AIMS Results.* Retrieved April 19, 2000 from the World Wide Web: http://www.ade.state.az.us.

Massachusetts Department of Education (1999). Massachusetts Comprehensive Assessment System: 1998 Technical Report (1999). Malden: MA.

Taylor, C.S. (2001). Washington Assessment of Student Learning: Grade 10, 1999 Technical Report. University of Washington: Olympia.