SCORING THE 2002 ASSESSMENT

The scoring process used for this assessment is a train-the-trainer model similar to that used in 1994 and in 1998. Table leaders were brought in seven days ahead of the scoring session to read, study, and discuss extensively a large sample of the work done by students and to study thoroughly the demands and expectations of the prompt and instrument designed to score the writing.

Through an extensive process of analysis and discussion, a set of anchor papers was selected by the table leaders to represent the descriptors for levels 1 to 5. These anchor papers, two per level, became the demonstrations of levels of writing described for each level. In addition, the leaders chose training papers and reliability papers to be used to ensure consistency among all scorers. The leaders then trained their own tables of six scorers using a common script and the anchors they selected as a committee as a whole.

During the scoring process, reliability papers were scored twice daily by all scorers, and the degree of consistency among tables and across the scoring floor was monitored. Any inconsistency was discussed, and the expectations were reviewed.

In all cases, scoring was done by teams of thoroughly trained scorers, who matched student responses with the criteria developed to measure student achievement. Using the general performance statements for each level, the specific descriptors for each level, and the anchor papers as illustration of the descriptions, scorers determined which description best identified the quality of student writing being examined. Rigorous procedures were in place to ensure the reliability of individual scorers and their consistency in applying scoring criteria. In addition, sophisticated management techniques developed over the history of SAIP assessments ensured a reliable and efficient process of managing student booklets and the data resulting from the scoring process. Statistical details about the scoring consistency will be provided in the forthcoming technical report.

Box 2

Technical Terms Used in this Section

Table leaders: educators from the jurisdictions brought in ahead of the scoring session to study student work, to select anchor, training, and reliability papers, and to learn to train and direct tables of six scorers each using a common script.

Anchor papers: examples of student work selected by consensus of the table leaders as the most representative of the description for a particular level in the criteria (see *Writing Assessment Criteria* sub-section above). Used repeatedly by scorers to confirm their understanding of the criteria during the scoring session.

Training papers: papers scored by all scorers for practice during training and at the start of each day, selected by table leaders to come to a common understanding in applying scoring criteria.

Reliability papers: common papers selected by table leaders to be scored by everyone simultaneously twice a day to check on the consistency of the scoring tables and to identify any need for retraining.

Exemplars: papers selected from the anchors to illustrate to the community what quality of student work was considered appropriate for each level of performance.

Prompt: description of the context and purpose for which the writing task is being done. In this case, it is the description of an environmental dilemma, and students are asked to *generate public awareness about the dilemma with a written piece.*

This report provides results for Canada as a whole, as well as those of individual jurisdictions. To facilitate understanding of the many graphs and charts that follow, this section includes a short note on interpreting the results.

In this report, most performance-by-level charts are based on **cumulative results** and actually show percentages of students **at or above** each level. The implication here is that students performing, for example, at level 5 have also satisfied the criteria for levels 1, 2, 3, and 4.

Differences

In this report the terms "difference" or "different," used in the context of performance levels and percentages, refer to a difference that is not due to chance. In a technical sense, they refer to a **statistically significant difference**. A difference is statistically different when there is no overlap of *confidence intervals* between the two measurements.

Confidence Intervals

In this assessment, the percentages calculated by the researchers are based on samples of students. Therefore, these are only estimates of the actual achievement students would have demonstrated had all students in the population taken the assessment. Because an estimate based on a sample is rarely exact, it is common practice to provide a range of percentages within which the actual achievement level might fall. This range of percentage values is called a **confidence interval**. It represents the high- and low-end points between which the actual achievement level should fall 95% of the time. In other words, one can be confident that the actual achievement level of all students would fall somewhere into the established range 19 times out of 20, if the assessment were repeated with different samples from the same student population.

In the charts in this report, confidence intervals are represented by the following symbol: ⊢ If the confidence intervals overlap, the differences are not statistically significant. It should be noted that the size of the confidence interval depends upon the size of the sample. In smaller jurisdictions, a large interval may indicate difficulties in achieving a large sample and does not reflect on the competency of the students who participated in the assessment.

Statistical vs. Educational Difference

Statistical significance is determined by mathematical formulas and considers issues such as sampling. It is a matter of interpretation as to whether a difference in results has educational significance. There are situations where a *statistically significant difference* may have little educational significance

---

Box 3

Statistical Comparisons

The performance of students in Canada (and within each jurisdiction) was compared by looking at the average scores for all students in each jurisdiction and at the distribution of these scores.

Because the available scores were based on samples of students from each jurisdiction, we cannot say with certainty that these scores are the same as those that would have been obtained had all 13-year-old and 16-year-old students been tested. We use a statistic called the *standard error* to express the degree of uncertainty in the scores for the sample compared with the population. Using the standard error, we can construct a *confidence interval*, which is a range of scores within which we can say, with a known probability (such as 95%), that the score for the full population is likely to fall. The 95% confidence interval used in this report represents a range of plus or minus about two standard errors around the average.

The following charts are intended as representations of numerical data and as such cannot always be interpreted with the same degree of precision as the actual numbers. This is particularly true for small percentages and small confidence intervals. For more precise data, please refer to the numerical tables in the appendix to this report, and to the forthcoming technical report.

(i.e., the difference is very small). There are also situations where a difference that is perceived to have educational significance may not in fact have statistical significance. For example, if one were to try to compare the 1994, 1998, and 2002 performances, the statistical differences would not be educationally significant in the light of changes to the test design. What may be educationally significant is the smaller gap between any one jurisdiction's level of student performance and the pan-Canadian performance in 2002 compared to 1998. Where applicable, these differences have been noted in the individual jurisdictional reports.

Comparisons Between Languages

Caution is advised when comparing achievement results based on assessment instruments prepared in different languages, despite the extensive efforts to ensure equivalence for the sake of equity and fairness for all students. Every language has unique features that are not readily equivalent. While the writing task, criterion descriptors, scoring scripts, and scoring process were highly equivalent in English and French, pedagogical and cultural differences related to differences in language structure and use render comparisons between languages inherently difficult.

SAMPLE CHART

The following chart is provided to help readers interpret the confidence intervals used in this report. For example, there is no significant difference between population L and populations A, C, E, F, H, I, J, and K, but there are significant differences between population L and populations B, D, and G because their confidence intervals do not overlap.

SAIP WRITING 2002: SAMPLE CHART
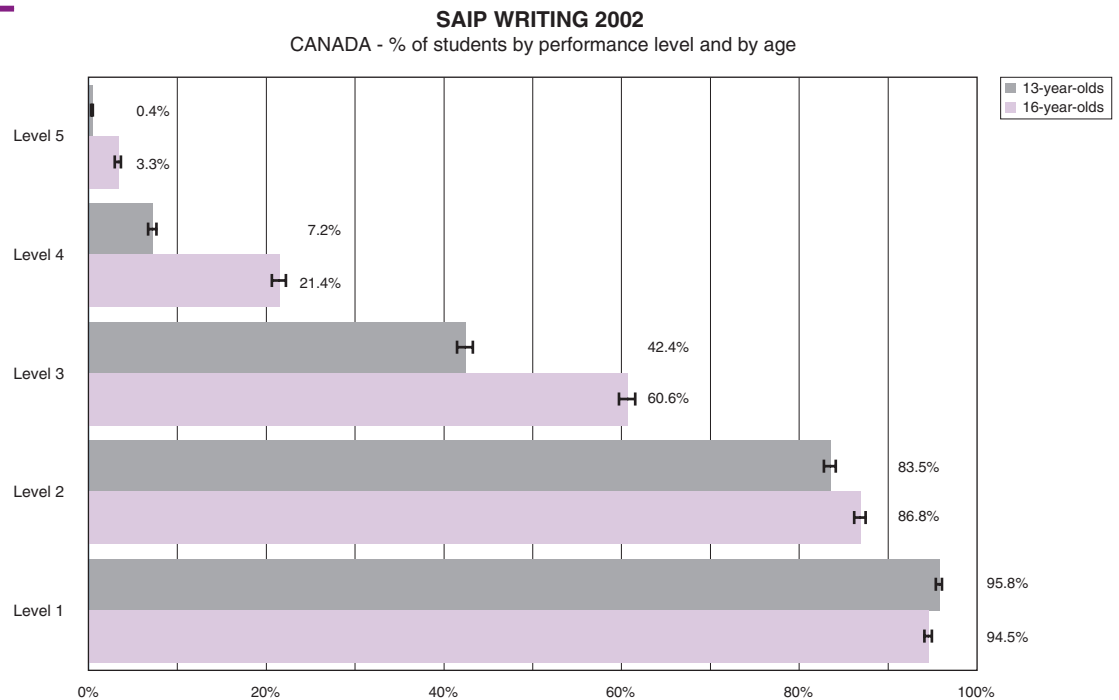Performance by population showing confidence intervals
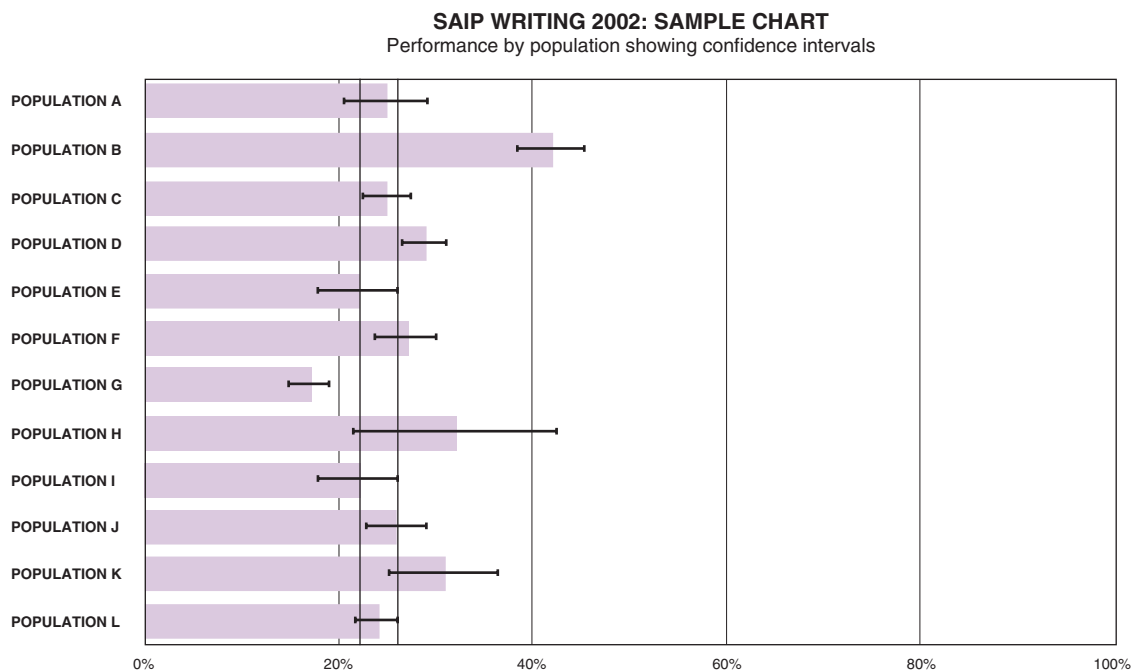
Introduction

In this section of the report, results are presented for Canada as a whole. Charts C1 to C3 compare overall Canadian results combining performance from all jurisdictions and both languages. Charts C4 and C5 compare expectations-setting results.

- Chart C1: % of students by performance level and by age
- Chart C2: % of 13-year-olds by performance level and by gender
- Chart C3: % of 16-year-olds by performance level and by gender
- Chart C4: Results and Expectations — % of 13-year-olds by performance level
- Chart C5: Results and Expectations — % of 16-year-olds by performance level

CHART C1

SAIP WRITING 2002
CANADA - % of students by performance level and by age

Legend: 13-year-olds, 16-year-olds

Level	13-year-olds	16-year-olds
Level 5	0.4%	3.3%
Level 4	7.2%	21.4%
Level 3	42.4%	60.6%
Level 2	83.5%	86.8%
Level 1	95.8%	94.5%

The 1994 criteria descriptors, which are almost identical in the 2002 assessment, were developed after an extensive study of actual writing by 13-year-olds and 16-year-olds across Canada. As a result, the assumption of the SAIP assessments is that a majority of 13-year-olds will achieve level 2 and a majority of 16-year-olds will achieve level 3. In other words, a majority of 13-year-olds, given the demands and context of this assessment, will demonstrate an uncertain control of the elements of writing relative to purpose. Integration of some of the elements will be apparent but development will be inconsistently maintained, conveying a simple meaning (level 2). However, a majority of 16-year-olds will demonstrate a control of the elements of writing appropriate to purpose. Their writing will be generally integrated and the development maintained throughout with a clear perspective (level 3).

It is not surprising then that, according to chart C1, most of the older students write at or above level 3, compared with less than half of younger students. Furthermore, 21% of 16-year-old students write at level 4 or 5 compared with 7% of 13-year-old students. It is worth noting, however, that 42% of 13-year-olds do demonstrate a level 3 performance. These differences are expected in the light of age difference and opportunity to learn for 16-year-olds in comparison to 13-year-olds.

SAIP WRITING 2002

CANADA - % of 13-year-olds by performance level and by gender

- ■ Females
- ■ Males

Level 5 — 0.3% / 0.4%
Level 4 — 9.4% / 4.8%
Level 3 — 50.0% / 34.3%
Level 2 — 88.5% / 78.1%
Level 1 — 97.3% / 94.2%

(0% — 20% — 40% — 60% — 80% — 100%)

SAIP WRITING 2002

CANADA - % of 16-year-olds by performance level and by gender

- ■ Females
- ■ Males

Level 5 — 3.9% / 2.8%
Level 4 — 26.0% / 17.3%
Level 3 — 69.4% / 52.9%
Level 2 — 91.9% / 82.6%
Level 1 — 96.9% / 93.0%

(0% — 20% — 40% — 60% — 80% — 100%)
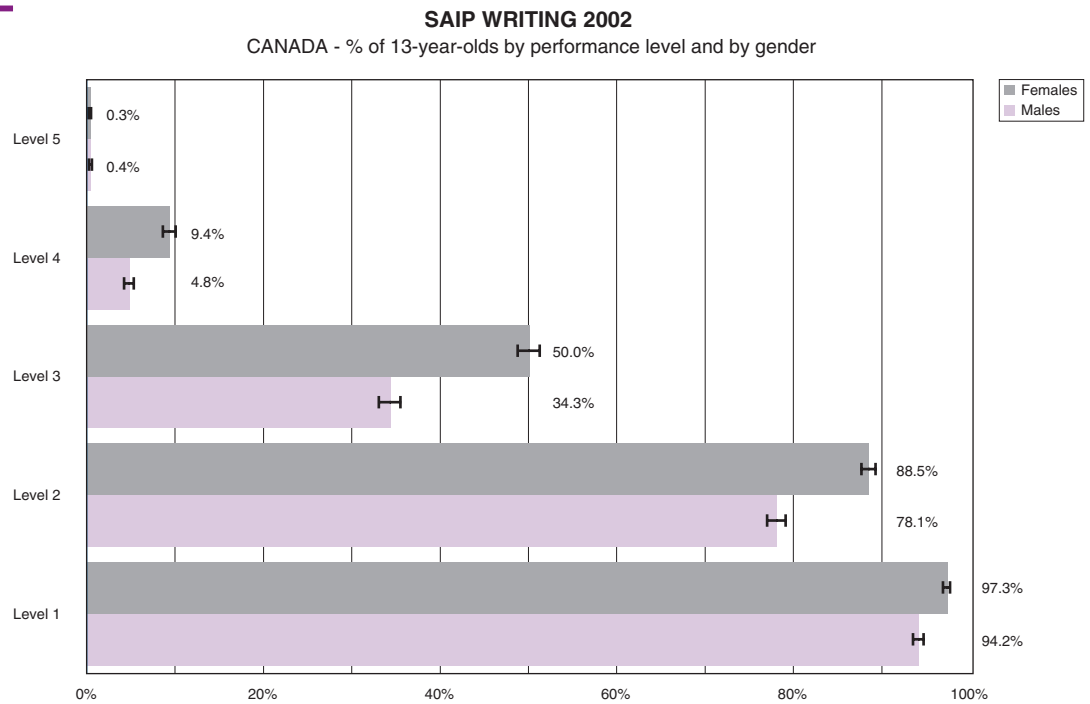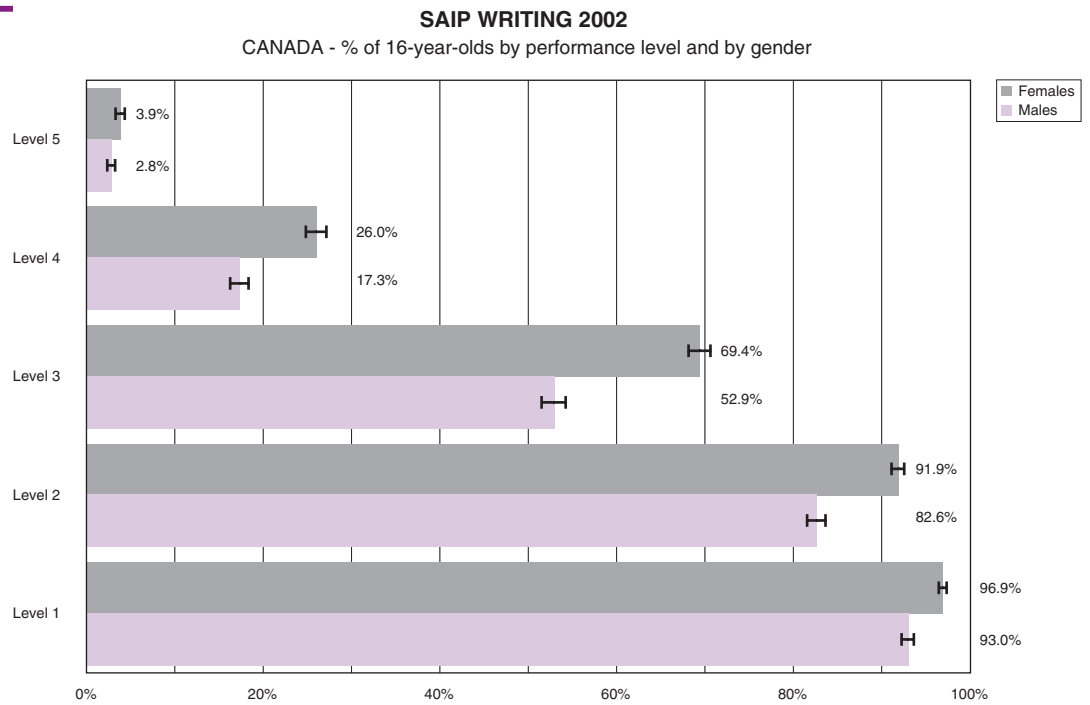
The 2002 SAIP assessment confirms what is now recognized as an international phenomenon in many cultures and languages, namely that girls demonstrate writing skills at a significantly higher level than boys. These charts indicate that the differences between girls and boys at each performance level are similar in both age groups. For 13-year-olds, the difference at level 2 is 10%, at level 3 around 16%, and at level 4, 5%. Among 16-year-olds, the difference at level 2 is 9%, at level 3, around 16%, and at level 4, 9%. The differences found in the SAIP Writing III assessment are consistent with those found among Canadian youths in the OECD PISA study of reading skills.[2]

_____

[2] Knowledge and Skills for Life: First results from PISA 2000. Executive Summary. OECD Publication.

An important question that must be asked for any assessment is one of expectations. "What percentage of Canadian students should achieve at or above each of the five performance levels, as illustrated by the framework and criteria and by the writing task?" The answer to this question must come not only from educators, but also from the broadest possible spectrum of Canadians.

To assist with the interpretation of SAIP assessments, CMEC regularly convenes pan-Canadian panels of educators and non-educators to examine the framework and criteria and to review the assessment instruments and scoring procedures. For the Writing III Assessment, panellists attended one of the three sessions held in Atlantic, Central, and Western Canada during October 2002. This panel consisted of teachers, students, parents, university academics and curriculum specialists, Aboriginal teacher trainers, business and industry leaders, community leaders, and members of national organizations with an interest in language education. The panel featured representatives from across Canada.

The approximately 100-member panel reviewed all assessment instruments, scoring procedures, and actual student results to determine the percentage of 13- and 16-year-old students who should achieve each of the five performance levels. Full and open disclosure was provided to panellists of any information pertinent to the assessment, including sampling procedures and data regarding the varying opportunities students across the country have to develop writing skills.

A collaborative process was used to define pan-Canadian expectations for student achievement in writing. Specifically, participants were asked to answer independently the question "What percentage of Canadian students should achieve at or above each of the five performance levels, as illustrated by the framework and criteria and by the writing task?" Panellists' answers to that question were collected to determine the desired Canadian student performance and to help interpret how students should do in comparison with actual results.

Box 4

How well did Canadian students REALLY do?

To ensure that the design and the results of SAIP assessments are really representative of the expectations that Canadians have for their students and schools, a broadly based panel is gathered from across Canada of both educators and representatives from business and the general public.

In three geographically based sessions, members examine all of the testing materials and share their expectations of how well Canadian students should perform.

Results of these sessions are then compared with the actual results and reported in the public report.

The following two charts show that the expectations-setting panel is generally pleased with the performance of Canadian students in writing. Generally, students in both age groups perform at expected levels. At all levels but level 3 among 16-year-olds, student performance falls within the range expected of them. Notably, at the highest levels (4 and 5), the performance is closely aligned to expectations.

CHART C4

SAIP WRITING 2002
CANADA - Results and Expectations
% of 13-year-olds by performance level

Legend:
- ■ Results
- ○ Median
- ■ Interquartile Range

Level 5: 0.4% / 3.0%
Level 4: 7.2% / 10.0%
Level 3: 42.4% / 45.0%
Level 2: 83.5% / 83.0%
Level 1: 95.8% / 95.0%

CHART C5

SAIP WRITING 2002
CANADA - Results and Expectations
% of 16-year-olds by performance level

Legend:
- ■ Results
- ○ Median
- ■ Interquartile Range

Level 5: 3.3% / 5.0%
Level 4: 21.4% / 25.0%
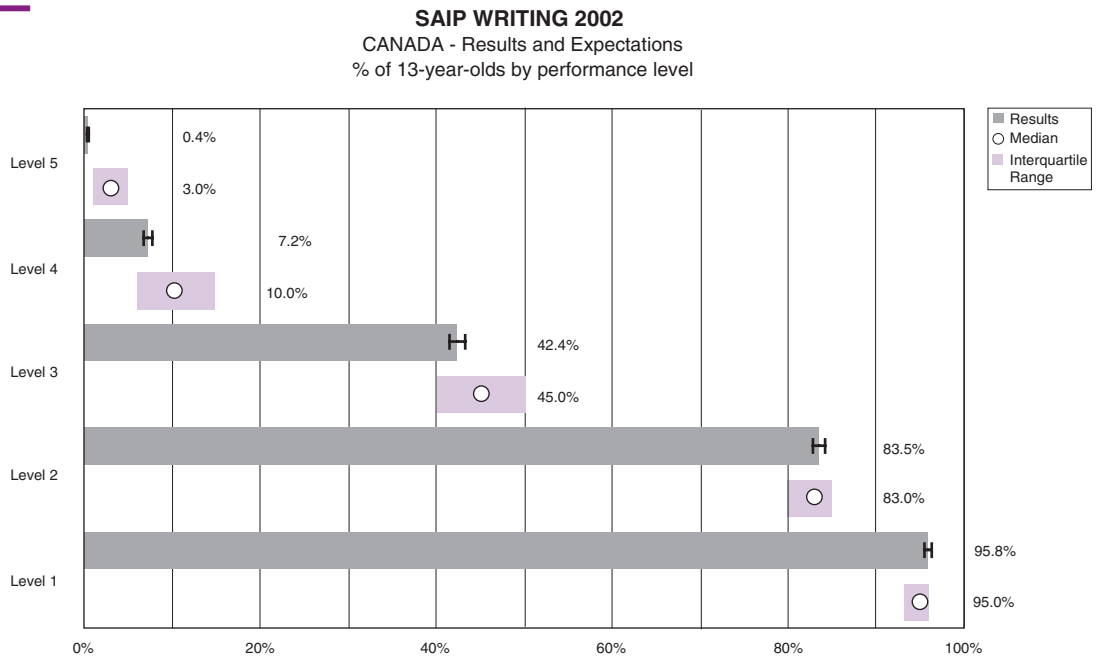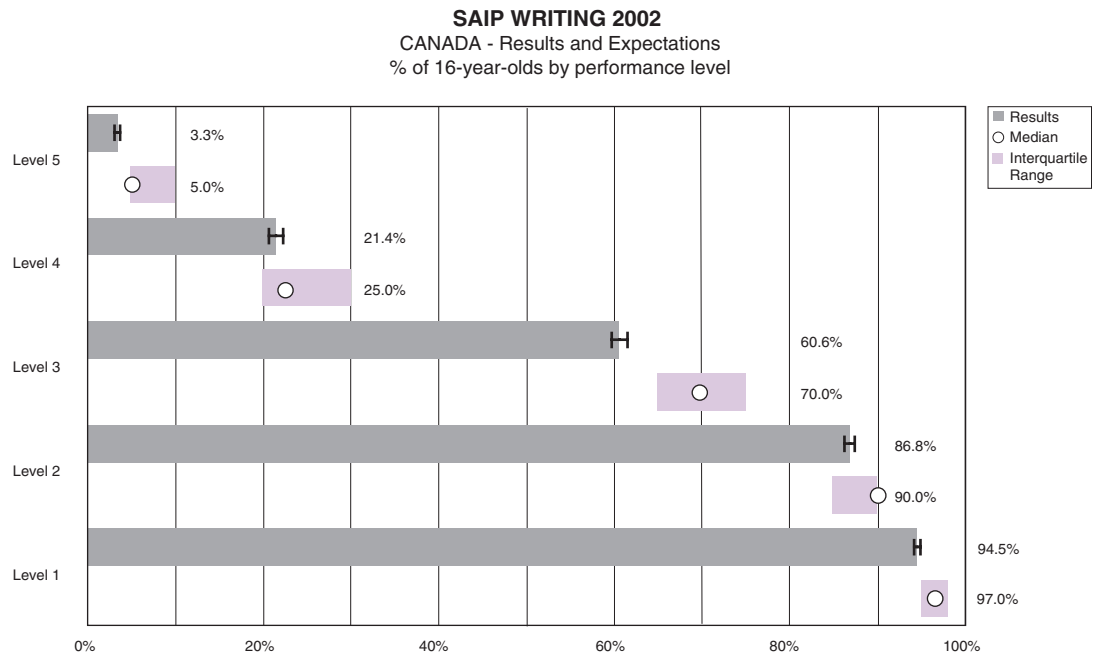Level 3: 60.6% / 70.0%
Level 2: 86.8% / 90.0%
Level 1: 94.5% / 97.0%

RESULTS FOR THE JURISDICTIONS

This section of the report presents the charts providing a comparative view of all jurisdictions as well as results by single jurisdictions. Results are shown for each participating jurisdiction. On specific jurisdictional results, comparisons are made to the Canadian results by language. That is, the English jurisdictions are compared to the Canadian English average, and the French ones to the Canadian French average.

The table below presents those participating jurisdictions, which showed a proportion of students similar to, or exceeding, the proportion of students meeting level 2 for 13-year-olds and level 3 for 16-year-olds.

OVERVIEW OF ACHIEVEMENT BY LEVEL

Table 3
Jurisdictions[1] performing better than or about the same as Canada[2]
13-year-old students at level 2

Canada (E)[3] (82.4% achieved level 2 or better.)	Alberta British Columbia Manitoba (E) Ontario (E) Quebec (E)
Canada (F) (87.3% achieved level 2 or better.)	Quebec (F)

16-year-old students at level 3	
Canada (E) (57.7% achieved level 3 or better.)	Alberta British Columbia Manitoba (E) New Brunswick (E) Newfoundland and Labrador Ontario (E) Quebec (E) Saskatchewan
Canada (F) (71.7% achieved level 3 or better.)	Quebec (F)

[1] Jurisdictions appear in alphabetical order.
[2] Differences in scores are statistically significant only when confidence intervals DO NOT overlap.
[3] (E) = English population (F) = French population

DISTRIBUTION OF PERFORMANCE LEVELS

The following charts present the percentage of students at each achievement level for all jurisdictions plus Canada. The data shown is an overview and displays the distribution of students at each achievement level.

The results do vary from jurisdiction to jurisdiction. In some cases achievement is significantly different from another jurisdiction, from the results within English or French, or from Canada as a whole.

Box 5
Please note that the charts that follow are **not** cumulative, that is, the bars represent the actual percentage of students at a particular level, rather than those who have achieved a particular level and above.

As before, percentages are based on samples of students. For all populations, the performances are only estimates of the actual achievement students would have demonstrated if all of the students in the population had taken the assessment.

CHART C6

SAIP WRITING 2002
Distribution of performance levels of 13-year-olds: Jurisdictions and Canada

	Below 1	Level 1	Level 2	Level 3	Level 4	Level 5
BC	5%	15%	44%	31%	6%	0%
AB	6%	11%	42%	34%	6%	0%
SK	4%	21%	43%	27%	5%	0%
MB(E)	5%	12%	39%	36%	7%	1%
MB(F)	4%	21%	47%	26%	2%	0%
ON(E)	3%	11%	42%	36%	7%	0%
ON(F)	8%	13%	49%	27%	3%	1%
QC(E)	4%	17%	40%	32%	6%	0%
QC(F)	3%	8%	35%	44%	9%	1%
NB(E)	6%	17%	43%	30%	4%	0%
NB(F)	4%	18%	47%	28%	3%	0%
NS(E)	7%	18%	46%	26%	3%	1%
NS(F)	3%	25%	50%	20%	2%	0%
PE	5%	17%	44%	30%	3%	0%
NL	8%	17%	42%	28%	5%	0%
YT	6%	26%	37%	26%	4%	0%
NT	11%	31%	36%	19%	2%	0%
CAN(E)	4%	13%	42%	33%	6%	0%
CAN(F)	4%	9%	37%	42%	8%	1%
CAN	4%	12%	41%	35%	7%	0%
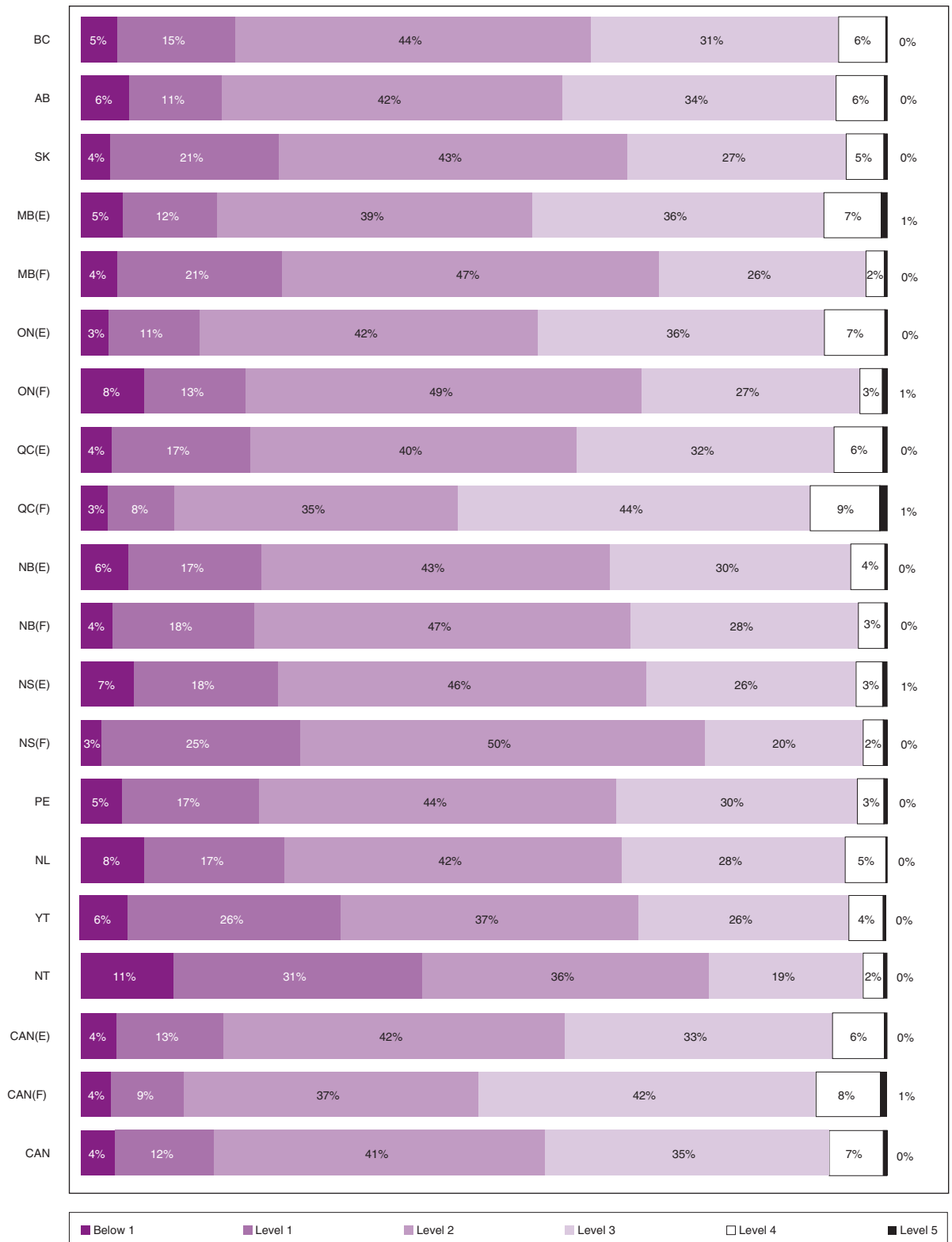
SAIP WRITING 2002
Distribution of performance levels of 16-year-olds: Jurisdictions and Canada

Jurisdiction	Below 1	Level 1	Level 2	Level 3	Level 4	Level 5
BC	5%	8%	31%	40%	14%	3%
AB	3%	7%	31%	38%	19%	3%
SK	3%	9%	31%	42%	13%	2%
MB(E)	5%	6%	29%	40%	17%	3%
MB(F)	12%	11%	35%	33%	8%	1%
ON(E)	6%	9%	28%	39%	16%	2%
ON(F)	8%	13%	34%	34%	9%	1%
QC(E)	3%	5%	25%	45%	19%	3%
QC(F)	7%	4%	14%	37%	29%	8%
NB(E)	5%	8%	28%	42%	15%	2%
NB(F)	4%	10%	29%	44%	11%	2%
NS(E)	6%	8%	33%	39%	12%	2%
NS(F)	1%	12%	45%	33%	9%	1%
PE	9%	11%	28%	37%	12%	2%
NL	6%	6%	30%	44%	13%	1%
YT	11%	15%	23%	36%	13%	1%
NT	10%	16%	31%	29%	12%	2%
CAN(E)	5%	8%	29%	40%	16%	2%
CAN(F)	7%	5%	16%	37%	27%	8%
CAN	5%	8%	26%	39%	18%	3%

Legend: ■ Below 1   ■ Level 1   ■ Level 2   ■ Level 3   □ Level 4   ■ Level 5