A DECADE OF SAIP

WHAT HAVE WE LEARNED; WHAT DO WE NEED TO KNOW?

Robert K. Crocker
Faculty of Education
Memorial University of Newfoundland
St. John's, Newfoundland
A1B 3X8

Presented at the

2001 Conference on Empirical Issues in

Canadian Education

Ottawa

November, 2001

Draft: Not for Citation

Background

The School Achievement Indicators Program (SAIP) was initiated by the Council of Ministers of Education, Canada in the early 1990s, in response to concerns about public accountability in education and particularly about the productivity of education systems in Canada. The fundamental goal of SAIP is to answer the question: "How well are Canadian students doing in the core school subjects of reading and writing, mathematics, and science?" Since the first assessment conducted in 1993, SAIP has gone through two complete cycles in mathematics, reading and writing and science, and is now entering its third cycle with mathematics in 2001 and writing scheduled for 2002.

The same tests have been administered to samples of 13- and 16-year old students, to provide an indicator of growth over the late middle school years. Following a 1999 enhancement, comparative data are also now becoming available on a wide range of student background variables as well as school and classroom conditions. Since the second cycle, data have also been available on performance expectations, based on the work of expert panels. Although originally conceived as a comprehensive educational indicators program, SAIP has evolved essentially into a comparative achievement study. A total of 18 populations have been defined using jurisdictional and language divisions (13 jurisdictions and two languages within five of these). Public reports have focused on comparing achievement levels across educational jurisdictions and across official language groups within some jurisdictions.

SAIP is designed to yield data at the national and provincial levels, but not at the individual student, school, or school district levels. Sampling and administration procedures have been developed with this goal in mind. Minimum error rates have been specified, and samples have been chosen to met these specifications. The data have typically been reported by province/territory, in the form of proportions of students at or better than each of five defined levels of achievement. Designation of these levels, and the use of expectations-setting procedures, are intended to allow the results to be interpreted in criterion-referenced terms, although the comparative approach implies that most interpretations are more normative in nature.

This paper examines the yield of SAIP over the first two cycles and the potential for added value from the program. Emphasis is placed on differences across jurisdictions and language groups and on the relationship between observed and expected performance. Using the 1999 science assessment, important differences in students, schools and teaching practices are also summarized. Finally, some

research questions are identified to which answers might be sought using SAIP and the many other large scale data bases that are now becoming available.

Are there Patterns of Differences among Jurisdictions?

The approach taken in presenting the SAIP results clearly indicates the interest in answering the basic question of how well students are doing, not only for the country as a whole, but for each of the jurisdictions. In the absence of a clear criterion or more correctly, because criterion-based statements are difficult to interpret, normative statements about how the jurisdictions are doing relative to the Canadian average and, less explicitly, to each other are inherent in the design of all of the reports.

The six assessments conducted to date, with multiple measures in most assessments, yield a total of 20 different data points for most of the SAIP populations. Comparing the performance of each population with the Canadian composite proportion on all measures yields a total of 296 available comparisons Many more can be generated using pairwise comparisons of jurisdictions.

It now seems reasonable to ask questions about trends in the results. For example, we might ask if there is symmetry or skewness in high and low performance, whether some populations show consistently high or low performance, whether there are trends towards improvement on the part of some populations and whether particular populations are relatively better in some subjects than in others.

Because the proportions used to express the results are directly comparable only within a particular measure, it is difficult to find a clear way of examining trends. One attempt to do this appears in the 1999 Pan-Canadian Educational Indicators Report (Statistics Canada and CMEC, 1999). This report gave a chart indicating, for each measure, whether a particular population was significantly higher or lower than or not significantly different from the Canadian level, using the proportions for Level 2 for 13-year-olds and level 3 for 16-year-olds. Counting the frequency of the three categories across all assessments gives a sort of “box score”of performance over the whole set of measures. This summary, with the 1999 science results added, is given in Figure 1.

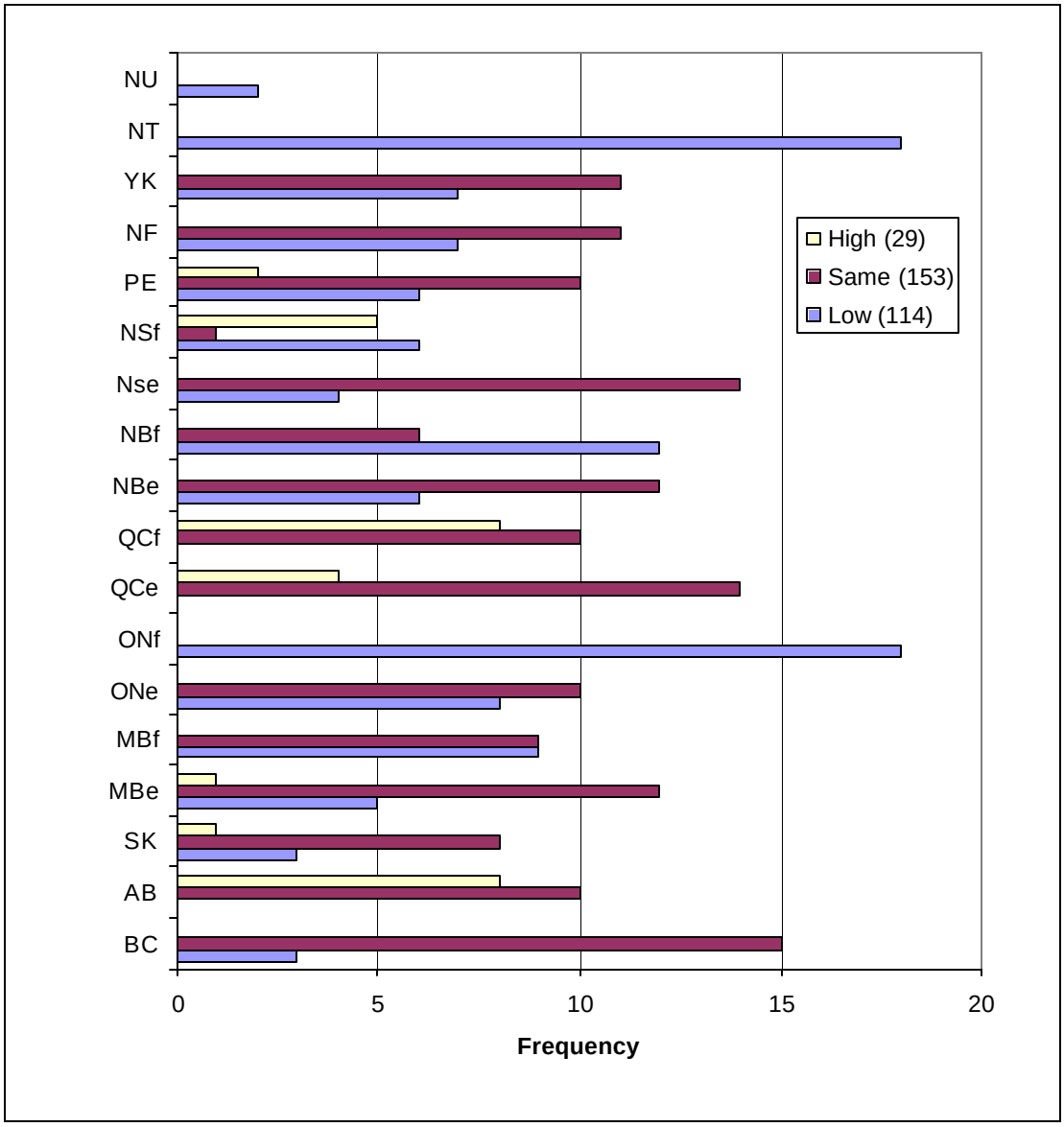

Figure 1
Jurisdiction Performance Relative to Canada
SAIP Cycles 1 and 2

This chart indicates, first, that significant differences occur much more often than would be expected by chance. In fact, individual populations are different from the Canadian composite proportion about half the time.[1] Generally, this tells us that performance is not uniform across the country. What is more striking is the pattern of highs and lows. Many more statistically significant lows than highs appear. This skewness suggests that it is difficult for a jurisdiction to bring its achievement level much above the national level. One possibility is that this is an artifact of differences in jurisdiction size. However, this phenomenon would place Ontario and Quebec closer to the composite than other jurisdictions, which is clearly not the case here. A more likely possibility is that this represents a "ceiling" effect, in which some jurisdictions are approaching a level that would be difficult to exceed. This hypothesis tends to be supported by the fact that the proportions reaching levels higher than those used as the benchmarks for the two age groups drop off substantially in all cases.

The question of consistency in performance can be examined using this chart. It is clear that some jurisdictions, notably British Columbia, Nova Scotia (English) and Quebec (English) tend to be near the composite on almost all assessments. Others, particularly, the Northwest Territories, Ontario (French) and New Brunswick (French) are consistently below the national average. The SAIP results have often been interpreted as showing low performance for francophone populations outside of Quebec. These comparisons indicate that this pattern is consistent only for francophones in Ontario and New Brunswick. Nova Scotia (French) results have been highly variable from high to low, while Manitoba (French) has varied from the same to low. While no jurisdiction has been consistently high, Alberta and the two Quebec populations show the best performance overall, accounting for most of the highs and having no lows.

The pattern of highs and lows is clearly subject-related, as indicated in Figure 2. Reading and writing performance is more likely than either science or mathematics to be the same across jurisdictions. There is a strong possibility that this is a function of test reliability or of restriction of range, as the proportions reaching the target levels for reading and writing are substantially higher than those for science and mathematics. There is a greater tendency for mathematics to be low than for other subjects. Breakdowns by jurisdiction are more difficult to interpret here, because of low cell frequencies. However, it is clear that high performance in science and mathematics are related, with the same few jurisdictions accounting for all of the highs in both areas.

[1]The Canadian composite proportion is weighted to account for different population sizes in the jurisdictions.

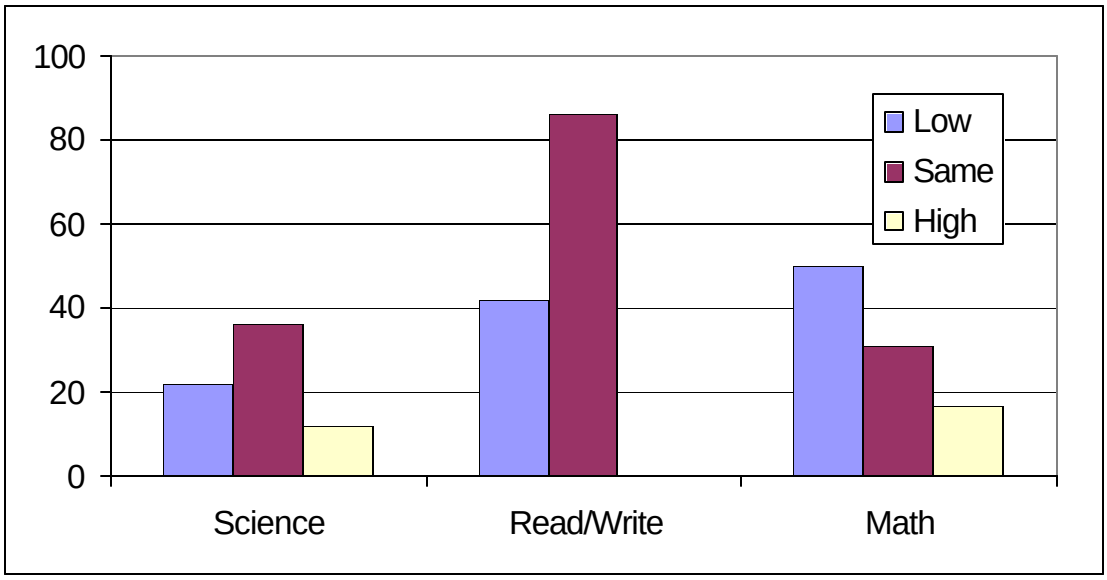

Figure 2
Performance Differences by Subject: Canada

Has Performance Improved Over Time?

One of the obvious purposes that can be served by large scale assessment is to encourage improvement in performance, especially on the part of low-performing jurisdictions. Although two cycles are insufficient to establish clear trends, it is nevertheless interesting to highlight what has happened over the complete second cycle. This is especially so since there are indications that Canada has improved its ranking in international comparative studies in recent years. This raises the question of whether we have seen a real improvement in Canada or if the relative improvement is at the expense of decline in some other countries. Again, the comparisons made are for reference levels 2 and 3 for 13- and 16-year-olds respectively.

Figure 3 compares performance across the two cycles for Canada as a whole.[2] mathematics showed a significant decline in performance among 13-year-olds in both content and problem-solving but an increase for 16-year-olds for problem-

[2]While confidence intervals are shown on the graph, a large sample test for the difference between proportions, rather than non-overlapping confidence intervals, was used to determine statistical significance.

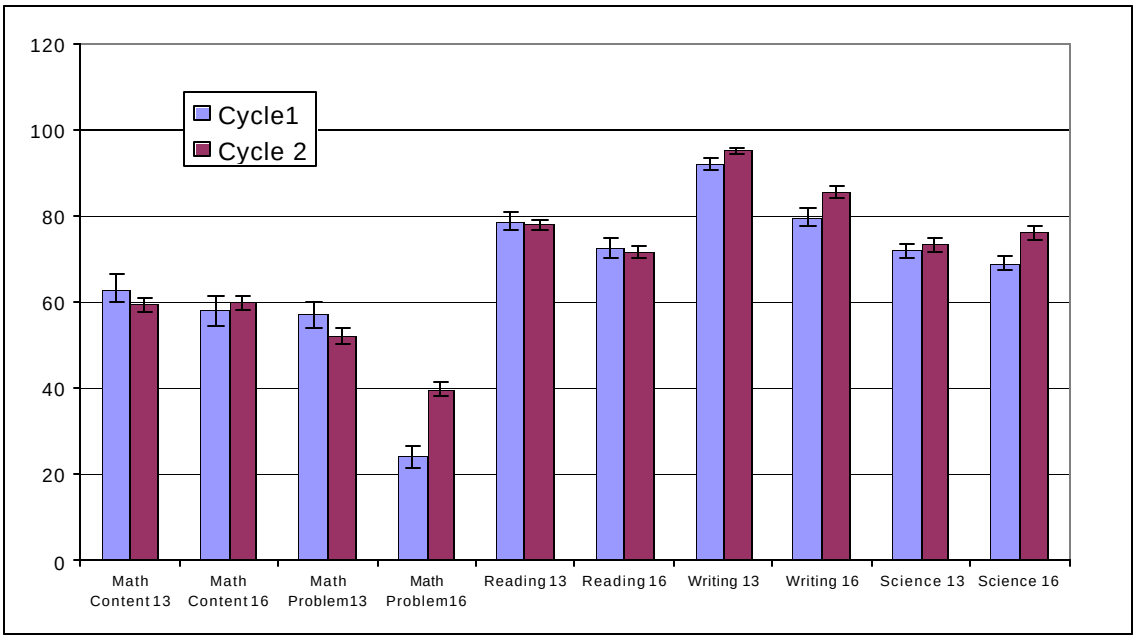

Figure 3
Performance Changes from Cycle 1 to Cycle 2

solving. No change was registered for reading, but writing improved for both age groups. Science showed improvement for 16-year-olds but not for 13-year-olds. Most of the jurisdictional changes were in the same direction as the change for Canada as a whole. However, these differences did not generally reach statistical significance because of the larger standard errors within jurisdictions.

On balance, the limited information available suggests a marginal improvement, with the only exception to this trend being the mathematics performance of 13-year-olds. With the results of the 2001 mathematics assessment due for release soon, it will be interesting to look for any longer term trend in mathematics performance.

Are Achievement Levels Meeting Expectations?

It is difficult to answer the original question of "How well are Canadian students doing?" using comparative results alone, because the comparative answer always implies the further question "Relative to what?" Beginning with the 1996 science assessment, an expectations-setting procedure has been used in the SAIP assessments. The procedure has involved convening regional panels of content experts, teachers and members of the public. Following a briefing on SAIP

procedures and the results of the latest assessment in the subject or interest, panellists were asked to respond to the question “What percentage of Canadian students should achieve at each of the five performance levels as illustrated by the Framework and Criteria and by the questions asked?”

Possible gaps between observed and expected results are better illustrated by examining the expectations at higher levels than those used in previous comparisons. For this reason, both levels 2 and 3 are examined for 13-year-olds and levels 3 and 4 are examined for 16-year-olds. Results of these comparisons, for the four assessments for which expectations are available, are given in Figures 4 to 7. The actual results are given as proportions of students in the sample at or above the level. The expected results are median proportions given by the expectations-setting panels.

Figure 4 clearly indicates that a large gap exists between results and expectations for mathematics for all comparisons made. Reading and writing tend to show the opposite effect, with performance exceeding expectations for writing and being fairly closely matched for reading. Two sets of comparisons are available for science. The 1996 assessment showed expectations exceeding performance on most comparisons. The match was closer in 1999, with an overall increase in performance, accompanied by some small shifts in expectations.

It is important to note that there is some difficulty in interpreting these results because the inter-quartile ranges for expectations given in the SAIP reports, are not directly comparable to the standard errors given for the actual results. Nevertheless, it is clear that the differences reported for mathematics are far in excess of what could be attributed to sampling errors. Indeed, in almost all cases, the actual proportions are substantially below the lower limit of the inter-quartile ranges. There is little doubt that public and professional reviewers expect students to do much better in mathematics than the results reveal. Whether this problem lies in unrealistic expectations or unsatisfactory performance is somewhat less clear.

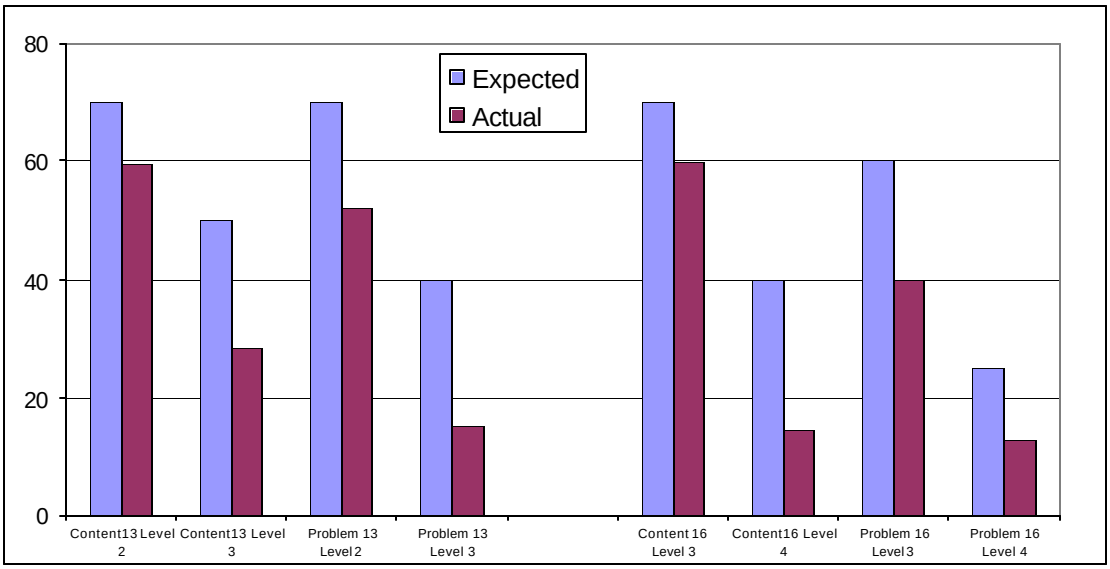

Figure 4
Results and Expectations: Mathematics, 1997

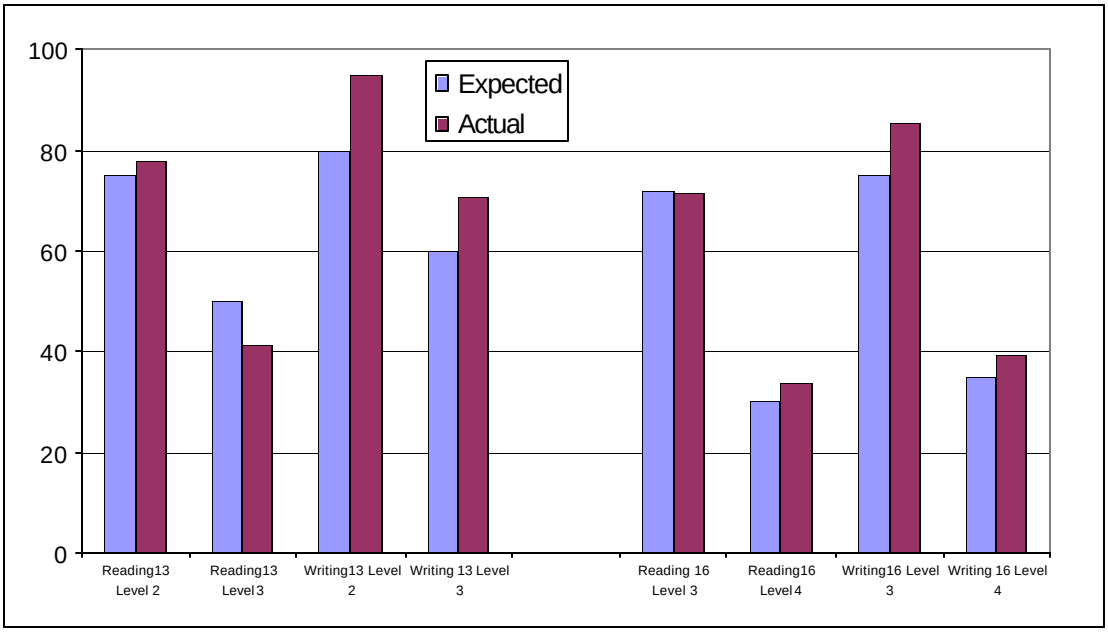

Figure 5
Results and Expectations: Reading and Writing, 1998

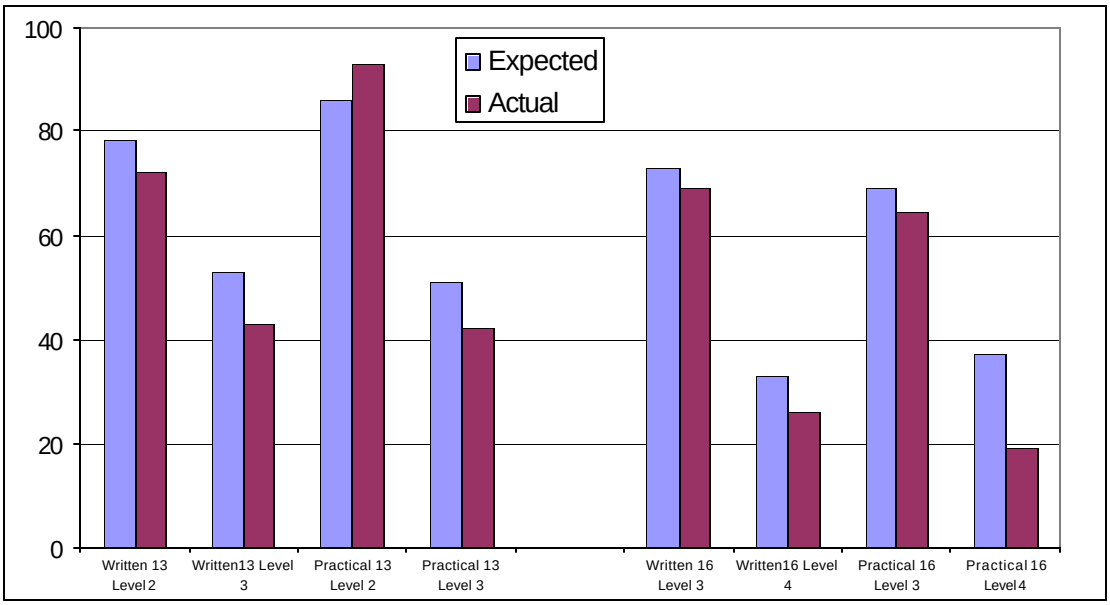

Figure 6
Results and Expectations: Science, 1996

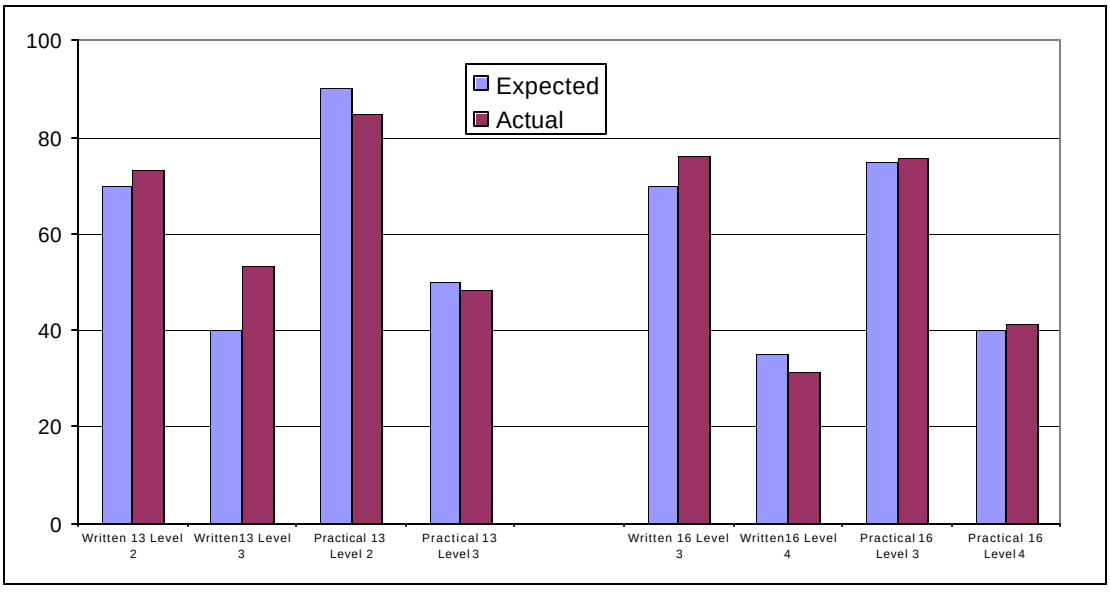

Figure 7
Results and Expectations: Science, 1999

The Context of Learning

Since the beginning, SAIP has used brief student questionnaires to capture some aspects of student characteristics and attitudes, and a few exploratory attempts have been made to examine relationships between these variables and achievement levels. For example, the 1993 mathematics assessment found some positive relationships between achievement and such variables as homework, calculator use and liking for mathematics but none between achievement and television-watching or computer use. Both the 1994 and 1997 reading and writing assessments found small positive relationships between achievement and activities related to the are area of assessment, such as liking for reading, books in the home, reading for pleasure and editing and revising writing. The 1997 science assessment found positive relationships between achievement and belief in hard work and studying at home and confidence in their ability to do science work, but a negative relationship between achievement and teamwork and achievement and frequency of laboratory activities.

A much more comprehensive set of student, teacher and school questionnaires was developed for use with the 1999 science assessment. The conceptual framework for the questionnaires was based on an elaboration of the **inputprocess→outcome** model commonly used in educational indicator studies. The specific version of the model was developed from work of Wang, Haertel and Walberg (1993). Using a synthesis of the review literature on factors associated with achievement, these authors identified six major classes of variables on an approximate proximal→distal continuum, ranging from broad policies to day to day classroom practices. Results of this synthesis tended to support the hypothesis that more proximal variables are more closely associated with achievement than more distal variables.

Modified versions of these questionnaires were used for 2001 mathematics and are under development for 2002 writing. The 1999 science report included a supplementary volume which presented descriptive/comparative results by jurisdiction. This approach was intended to highlight how the context of learning differs among the various educational jurisdictions, in a way that might be useful for policy analysis. No analysis of the links between the contextual variables and achievement has been conducted to date.

In general, the descriptive/comparative results showed substantial differences across jurisdictions on many important aspects of school functioning, teaching and learning and student attitudes and habits. A full review of the results is beyond the scope of this paper. The following highlights are intended to give a sense of the scope of the data gathered and of important pan-Canadian trends and jurisdictional and language differences.

Schools:

1. Average class sizes tend to be in the 20-25 range but vary substantially cross jurisdictions. Classes for 13 year olds tend to be slightly larger than for 16 year olds.

2. Levels of parental involvement in aspects of school life were generally reported by principals as low, with some variations across jurisdictions.

3. Community conditions, lack of parental support, student ability and home background were reported as more prevalent factors limiting instruction in francophone schools and those in the Territories than in other anglophone schools.

4. Shortage of science teachers and other specialists were more prevalent as limiting factors in Eastern and Territorial than Western Schools, with Quebec francophone schools reporting the lowest limitations.

5. Most schools reported having substantial numbers of computers, with a high ratio of up to date (defined as computers capable of running Windows-based programs and Web browsers) to total computers.

6. Schools are generally not streamed or ability grouped for 13 year old students. However, streaming is much more prevalent at the 16 year old level. Wide variations in the incidence of streaming are found across jurisdictions.

Teachers:

1. About 60% of teachers overall are female, with relatively small variations across jurisdictions. Most teachers tend to be in mid-career, with those in the Quebec anglophone system standing out as having substantially more experience and those in the Territories less experience than teachers elsewhere.

2. Almost all teachers hold university degrees, with the B.Ed. being most common. The proportion of teachers specialized in science, as evidenced by the B.Sc. degree or equivalent varies widely across jurisdictions. Relatively few teachers, less than 10% in most jurisdictions, hold master's degrees. A notable exception is the Quebec anglophone system, where more than 20% of teachers hold the advanced degree.

3. The level of involvement of teachers with parents is not very high and is characterized by wide variations across jurisdictions. Teachers in anglophone jurisdictions reported greater contact with parents than their francophone counterparts. The main source of contact is parent-teacher interviews.

4. There was general agreement between teacher and student reports on classroom activities. The most common activities during class sessions were reported as note-giving, showing students how to do problems, diagnosing individual student problems or weaknesses, students working alone on assigned work and the teacher working with individual students. The use of science books and magazines varies widely between language groups, with francophone teachers and students both reporting much less use than their anglophone counterparts.

5. The frequency of laboratory activities in science was variable across jurisdictions with a pattern of more laboratory activities in the three Western provinces and among both language groups in Ontario and Quebec.

6. Almost all teachers support the proposition that students need to work hard to do well in science but relatively few agree that success in science requires natural talent.

Students:

1. Generally more students in the Atlantic Region and Nunavut have parents with less than a high school education than those elsewhere.

2. Students have relatively high educational aspirations, with more than 90% indicating that they intend to continue their education beyond high school and little variation across jurisdictions.

3. About half of 16 year olds plan careers in fields related to science and technology, with relatively small variations across jurisdictions.

4. Strong language differences were apparent in student perceptions of how their teachers and parents view the importance of their doing well in school and science, with anglophone students having much more positive views on these matters than their francophone counterparts.

5. Generally less than half of the students reported doing one hour or more of science homework per week. Fewer francophone than anglophone students and fewer 13 year olds than 16 year olds reported the higher homework times.

6. Anglophone students tended to more positive views than francophone students on the quality of school life.

The most notable overall feature of these results is the substantial variation that occurs across jurisdictions. Even greater variation is evident among individual students, teachers and schools. From the perspective of future research that might be conducted with these data, such variation is desirable because this makes it more likely that any relationships that exist between policies, practices and attitudes and student achievement will be detected.

It is possible to identify from these results some general positive and negative features of schooling in Canada. For example, it is clear that the overall qualifications of teachers are high and that teacher views towards science are in accord with contemporary philosophical perspectives on the nature of science. Teachers and students appear to believe that science is important and that one can do well by working hard. Students have very high educational aspirations and substantial numbers plan careers in fields related to science. Generally, student attitudes towards school and science are positive.

On the more negative side, teachers and principals indicated that their levels of engagement with parents is not very high. (In the absence of parent data, it is not possible to confirm whether parents share that view). Despite the prevalence of computers both in school and at home, it appears that the computers is not commonly used as an instructional tool in science. Finally, of the various differences that have been reported between anglophone and francophone groups, the prevalence of more negative attitudes among francophone students compared to their anglophone counterparts is a source of some concern.

It worth cautioning here that no direct association should be inferred between any of the contextual variations and the achievement levels found for different jurisdictions. Such relationships are likely to be complex and multi-variate. Even at a descriptive level, however, the observed variations raise interesting questions for policy deliberation. For example, we might ask whether some of the variations observed are a function of provincial policies, characteristics of the local society or culture, or consequences of teacher training, school characteristics or other features of the system. We might also ask whether it is desirable to preserve variations or to bring the various systems closer together.

Policy Implications

Other then perhaps ignoring the results, or considering existing achievement levels to be either satisfactory or immutable, two main possibilities seem to exist for action based on the comparative results. First, the poorest performing jurisdictions might attempt to find ways to improve their relative standing. Plausible policy responses might be to overhaul curriculum, target resources to the area seen as deficient, establish specific improvement targets or simply point out specific areas of deficiency in the hope that teachers and schools will act to overcome these deficiencies. Over time, this would be expected to result in a convergence of results as the catch-up efforts take effect. Action of this sort is essentially designed to bring about greater equity in achievement levels, with the incidental result of improving the overall average.

The second possibility is that the results might engender a competitive race, with high performing jurisdictions taking action to ensure that their position is protected, will low performing jurisdictions attempt to catch up. The most direct consequence of this is overall improvement but no change in relative standing. However, other possibilities are also plausible. For example, it may prove to be more difficult to increase high achievement than low achievement, as would be the case is a ceiling effect is in place. In any case, this is clearly a scenario intended to improve overall achievement, without explicit attention to equity. Whether greater equity would result might depend on how difficult it is to increase achievement at higher relative to lower levels.

There are indications from provincial strategic plans that some jurisdictions have begun to establish achievement targets. In one particular case, these targets are actually expressed directly in terms of improving SAIP performance. In others, these are linked to provincial assessments. It is not clear from this, however, if the targeting is occurring differentially in low or high performing jurisdictions. Nevertheless, such targeting is a clear indicator of continuing concern with achievement and a desire to improve.

There is little doubt that the large gap between actual and expected performance in mathematics requires direct attention, especially if the upcoming 2001 report shows the same result. While research should be conducted to help determine if the problem lies with the expectations or the performance levels, it seems plausible to work initially from the premise that the problem is with achievement. This is because dismissing the gap as a flaw in expectations is can preclude any attempt to improve achievement.

Assuming that the goal of improving achievement is a legitimate one, a major difficulty in making policy inferences from these results is that descriptive/comparative analyses do not, in themselves, give any sense of what actions might be most

effective in achieving that goal. There is a temptation to make intuitive comparisons between achievement levels and contextual factors across the jurisdictions. However, it is clear from the history of research on achievement that no single factor has any dramatic effect and that focussing on one or two seemingly obvious actions can lead to costly policy errors.

Now that SAIP is beginning to generate a more comprehensive data base, it is useful to identify some broad questions which might be answered by a research program designed to take us beyond descriptive/comparative analysis. These same questions might also be pursued using other large scale data bases, such as TIMES or PISA, that have emerged in recent years.

1. What are the relative influences of student characteristics, school, and teacher/classroom variables on achievement?

2. Controlling for student characteristics, to what extent do broad policy-related variables such as school size, class size, teacher qualifications, use of resources, autonomy in decision-making, and use of time influence achievement?

3. More specifically, does the SAIP data support a hypothesized pattern of stronger influences for student background and classroom practices and weaker influences for policy related variables?

4. Is there an interaction between student background and school and classroom practices in influencing achievement?

5. What student attitudes and activities are associated with higher or lower levels of achievement?

6. Are specific classroom practices more highly associated with achievement than others?

7. Do certain school and classroom practices reduce inequalities in achievement between students of different socioeconomic and family backgrounds?

An argument can be made that the long-term value of SAIP and similar large scale achievement studies lies in the development of high quality data bases that can help us find answers to these questions. SAIP has not realized its potential as a policy tool for a number of reasons, including the political sensitivity of the program, the absence, until recently, of contextual data, and the lack of a comprehensive research program capable of addressing the many conceptual and technical problems inherent

in dealing with such large scale data sets.

The most obvious final point to make, from a researcher's perspective, is that policies need to be developed that would encourage the needed research. It is important to note that SAIP requires the investment of substantial amounts of public funds. Relative to the cost of data collection, the cost of conducting comprehensive analysis would be relatively small. A collaborative effort among CMEC, the funding agencies and researchers, with modest additional funding, would be the most appropriate way to ensure that SAIP yields greater value on the investment.

References

Council of Ministers of Education, Canada. SAIP Public Reports:

- **Mathematics, 1993,**
- **Reading and Writing, 1994,**
- **Science, 1996,**
- **Mathematics, 1997,**
- **Reading and Writing, 1998,**
- **Science, 1999.**
- **Science Learning: The Canadian Context, 1999.**