In November 2002, the Education Quality and Accountability Office (EQAO) embarked on a strategic initiative entitled “Ensuring Quality Assessments.” The overarching purposes of this initiative were to:

- Build a systematic and cyclical process for updating EQAO’s practices;
- Ensure assessment processes incorporate recent improvements in standards and exemplary practices for large-scale assessment; and
- Ensure that reporting meets user needs for accountability and improvement planning.

The initiative involved activities to:

- Ensure a reporting process that meets EQAO’s accountability mandate, address the information needs of school board, schools, and parents, and are timely in terms of school and board needs;
- Review, refine and/or affirm the processes for each assessment and seek congruence, where feasible, across assessments in blueprints, construct validity, marking and analysis, item development, field testing, and year-to-year comparability; and
- Streamline administrative processes related to preparation, administration, and tracking of the assessments.

These activities were organised into a four-part review process: (1) research and study, (2) consultation, (3) analysis and synthesis, and (4) implementation.

*Research and study.* EQAO staff reviewed assessment programs in other jurisdictions, including the processes used for blueprints, item development, item banking, construct validity, year-to-year comparability, field testing, marking, and analysis. The processes used by other jurisdictions were compared to EQAO processes. EQAO staff also reviewed recent literature on standards and exemplary practices in large-scale assessment.

As part of the research and study phase of the Ensuring Quality Assessment initiative, EQAO established the “EQAO Assessment Review Project” by contract with a research team at the Ontario Institute for Studies in Education of the University of Toronto (OISE/UT). EQAO commissioned OISE/UT to form a research team, including a panel of international experts, and carry out an external review. The project was to consider current standards for large-scale assessment and to review, refine and/or reaffirm all processes associated with EQAO assessments for Grades 3, 6, and 9 and the Ontario Secondary School Literacy Test (OSSLT), including test design, assessment blueprints, scoring methods, reporting options, comparability, and reliability models.
This external review was conducted in six phases:

A. Initial review and recommendations concerning the assessment work currently in progress, which involved analysing the ongoing operational requirements and providing recommendations for implementation;

B. Analysis and technical audit of the current assessment systems to establish a clear scientific base for review of the design of testing and reporting systems;

C. Analysis of the substantive, psychometric, and statistical characteristics of the current assessment system to assess validity and reliability;

D. Investigation of fundamental scientific bases for assessments and evaluation of needs analysis carried out by EQAO during the consultation process;

E. Comparison, demonstration, simulation, and evaluation of design alternatives based on information EQAO already has on hand, input from project teams, and advisors; and

F. Recommendations for the design and testing along with the provision of training and assistance.

Consultation. To supplement and inform its research, EQAO held dialogue forums with the Assessment Advisory Committee and with representatives of stakeholder groups (principals, teachers, supervisory officers, board contacts, parents) regarding the nature and extent of EQAO reporting and stakeholders’ needs related to accountability, improvement planning, and staff development.

Analysis, synthesis, and implementation. The next steps in the Ensuring Quality Assessments initiative require the analysis and synthesis of the research and consultation results. Since November 2002, each assessment program has had at least one administration and materials for future administrations have continued to be developed. EQAO has monitored the research and consultation processes closely, when possible responding to recommendations as they emerged. As a result, EQAO is well on its way to implementing several of the recommended refinements to its assessments.

EQAO anticipates that the Ensuring Quality Assessments initiative will yield the following outcomes:

- Assessments that continue to meet emerging standards for large-scale assessment;
- Increased confidence by various stakeholders in the design of the assessments and in the validity and reliability of the results;
- Simplified administrative procedures for schools;
- School and board reports that meet EQAO’s accountability mandate, address the information needs of schools, boards and parents and contribute to improvement planning processes in schools and boards;
• An evolving philosophy and methodology of continuous improvement;
• Cyclical reviews that allow for stakeholder input through on-going consultation;
• The incorporation and implementation of changes in the field of large-scale assessment; and
• The incorporation of changing needs of stakeholders and policy makers into the design processes.

This report presents analyses, interpretations, and recommendations of the authors from the OISE/UT team and their consultants and do not necessarily reflect EQAO’s policies, evaluations, or plans. However, EQAO is pleased to present this report for wider discussion.

Marguerite Jackson, Chief Executive Officer, EQAO
Victoria Hemming, Chief Assessment Officer, EQAO
CONTENTS

PREFACE AND SUMMARY OF FINDINGS AND RECOMMENDATIONS ............................................. i
CONTRIBUTORS .................................................................................................................................. vi
ACKNOWLEDGEMENTS ................................................................................................................... ix

CHAPTER 1 INTRODUCTION .............................................................................................................. 1
  Ontario's Assessment Programs ........................................................................................................ 1
  This Review ........................................................................................................................................ 2
  Standards for Assessments .............................................................................................................. 2
  This Report ....................................................................................................................................... 3

CHAPTER 2 CLARIFYING THE PURPOSES OF THE ASSESSMENT PROGRAMS ....................... 5
  EQAO'S Mandate ............................................................................................................................. 6
  Purposes and Test Designs .............................................................................................................. 7
  Findings .......................................................................................................................................... 8
  Conclusion ...................................................................................................................................... 10

CHAPTER 3 DEFINING THE CONSTRUCTS AND SPECIFYING THE CURRICULUM CONNECTIONS ...................................................................................................................... 11
  Construct Definitions ....................................................................................................................... 12
  Curriculum Connections ............................................................................................................... 13
  Findings .......................................................................................................................................... 16
  Conclusion ...................................................................................................................................... 18

CHAPTER 4 DESIGNING THE ASSESSMENT PROGRAMS .............................................................. 19
  Design Considerations ................................................................................................................... 20
  Designs .......................................................................................................................................... 23
  Conclusion ...................................................................................................................................... 40

CHAPTER 5 DEVELOPING ITEMS ................................................................................................. 41
  The Process of Item Development ................................................................................................. 42
  Item Banking and Documentation Systems .................................................................................. 45
  Findings .......................................................................................................................................... 45
  Conclusion ...................................................................................................................................... 46

CHAPTER 6 GETTING FROM THE STUDENTS’ WORK TO SCORES ............................................. 49
  Marking the Student’s Responses ................................................................................................. 50
  Scaling the Students’ Responses ................................................................................................. 52
  Findings .......................................................................................................................................... 55
  Conclusion ...................................................................................................................................... 56
PREFACE AND SUMMARY OF FINDINGS AND RECOMMENDATIONS

The Education Quality and Accountability Office (EQAO) is an independent public agency charged with developing and administering tests in the elementary and secondary schools of Ontario and reporting the test results publicly. More generally, EQAO reports on the quality and effectiveness of elementary and secondary education and on the public accountability of the boards of education.

The first EQAO assessment program, by which we mean a large-scale educational testing and reporting process, was the reading/writing/mathematics assessment in Grade 3 during the 1996-97 school year. This was an extraordinary event: The assessment was carried out as a census, in that all Grade 3 students in Ontario participated; it involved performance tests that had students writing, reading, writing about reading, and doing mathematics activities; the testing was embedded in natural classroom settings and actually took two weeks of class time; the tests were developed as authentic representations of classroom activity and evaluation by teams of trained teachers and the holistic marking was carried out by trained teachers; and reports were returned to students and parents, teachers, schools, boards and the province within the school year. The outcomes were – in addition to individual and group test results linked to the Ontario school curriculum – professional capacity building in the curriculum and in the culture of evaluation. The assessment was bilingual, with equal development given to the parallel French and English tests. This was one of the largest comprehensive assessment programs anywhere in the world, considering the depth, seriousness, and sheer magnitude of student performance testing.

Over the years, EQAO has added assessments in Grade 6 for reading, writing and mathematics, in Grade 9 for mathematics, and in reading and writing literacy for the Ontario Secondary School Literacy Test (OSSLT). For each new assessment, EQAO has maintained the principles of high-quality performance testing and attended to consequential validity in a manner that places it at the forefront of large-scale assessment in Canada and around the world.

The educational context in which the EQAO assessments take place has not remained constant, however. The purposes the assessments are being asked to serve have changed and expanded. This is a common – and serious – problem for assessment programs.

Finding 1. The assessment programs have been asked to serve a myriad of purposes. EQAO has tried to accommodate these purposes. However, because it is impossible to design an assessment program to meet diverse purposes, the proliferation of purposes may compromise how well EQAO’s assessments can accomplish any one of them.

Recommendation 1. That EQAO continue to develop, in consultation with stakeholders, clear statements of the purposes of each assessment program, including the relative importance of each purpose.
The EQAO assessments have been aimed conscientiously at measuring and reporting student achievement with reference to the content and criteria specified in the Ontario educational curriculum and to the corresponding goals and practices for classroom instruction. This is especially important for an assessment program that aims to provide information for accountability and for improvement in student learning. Linking assessment to curriculum is a difficult job, first because the curriculum is a moving target and second because the curriculum specifications, which are intended for more general purposes, need to be analysed and transformed into assessment frameworks and eventually into tests.

Finding 2. EQAO has created framework documents based on the Ministry curriculum. However, these framework documents do not clearly define the construct or constructs each assessment is intended to measure, nor describe with sufficient detail the decisions that were made in interpreting the curriculum documents and developing the assessment frameworks.

Recommendation 2. That EQAO review, revise, and expand documents that define the constructs to be measured by the assessments and the connections between the assessments and the curriculum.

An assessment program design at a particular point in time depends on (1) the educational and political requirements as they are then understood; (2) the complexities of test development, administration, marking, scoring, and reporting; (3) the array of available technical methods for test analysis and comparison; and (4) costs and the available resources. EQAO has had to pick its way through shifting requirements, very complex measurement and technical methodologies, and limitations of time, materials, and funding. At each step, reasonable decisions have been made given the contexts and requirements. The present review and reappraisal of the assessment program is based on an analysis of the current and anticipated future contexts, resources, goals, and available methodologies.

Finding 3. Since 1996, the context for EQAO’s assessment programs has changed. Because each program was designed to fit a specific context, the designs may need to be revised to accommodate added purposes and refined construct definitions and curriculum connections.

Recommendation 3. That EQAO make decisions about the assessment programs’ designs after the purposes have been clarified, the constructs defined, and the curriculum connections specified. That EQAO plan to implement any changes over two to three years, to ensure that the quality of the assessment items and marking and scaling processes are maintained.

In particular, we recommend that EQAO consider the following design changes.

- For the Grade 3 and 6 assessments, move to a design that is modular rather than thematic. Reuse some modules across years to facilitate the equating of results.
• For the Grade 9 assessment, move to a two-part design with part to be marked by the mathematics classroom teacher and incorporated into student course grades and part to be marked centrally and provide information at the school, board, and provincial levels about trends in achievement in mathematics.

• For the OSSLT, if the construct definition for literacy supports a unitary literacy, move to a one-test, one-score, one-decision design; otherwise, if two constructs are clearly defined, separate it into two tests, one measuring reading and the other writing. Implement the change as soon as practicable.

Like all testing agencies, EQAO has had to face tradeoffs between operating costs, detail in scoring and reporting, student and teacher time, and speed in processing and return of results. Given EQAO’s commitment to performance assessment, which requires lots of time to administer and to mark, the decisions have been difficult. At each step in the development of the assessment programs, EQAO has made sensible choices given the requirements and constraints. The reevaluation of purposes provides an opportunity to investigate the rebalancing of efficiency with timing and detail of measurement.

Finding 4. The current assessment programs require a considerable commitment of time by students and by schools. Because most of the student work must be marked by teachers during special summer sessions, the marking costs are high. The information currently reported is detailed, but may not be optimal for its expanded uses. It may be that, with careful redesign, the tests can be shortened while providing information better targeted to the needs of schools and the Ministry.

Recommendation 4. That EQAO, once the purposes of the assessments have been clarified, tailor to these purposes its requirements concerning comprehensiveness and detail of the measurement, and concerning accuracy and comparability of scores and reports at the individual, school, board and provincial levels. That these requirements, along with considerations of cost and time, determine which design is selected.

EQAO is now approaching ten years of operation. It has carried out about twenty large-scale assessment surveys, including the regular Grade 3, 6, and 9 assessments, and the OSSLT. Thousands of test items have been written, edited, and validated by EQAO using internal and contracted staff, and many of the items have been very good. In a testing program, item quality is essential. Items must have a clear linkage to the curriculum content, be unbiased for all students, and be technically accurate, and must in aggregate cover the subject area comprehensively. EQAO has generally been able to meet the requirements for number and quality of items. However, a continuing testing program, with shifting and expanding demands, always needs more and better items, and the only internal answer is to systematise development. (An external answer is to buy items, but that is difficult when the testing is to be curriculum-specific and bilingual.)
Finding 5. While many of the items on EQAO’s assessments are of good quality, the processes by which items are developed, reviewed, and tried out allows some items of lesser quality to appear on the tests.

Recommendation 5. That EQAO establish and follow a standard process for the development and verification of items across all assessment programs. That EQAO create and maintain a general-purpose item banking and documentation system to support this process.

The EQAO assessments were each developed in response to specific circumstances and goals. For example, the original Grade 3 assessment represented an attempt to introduce new approaches to classroom evaluation and to interpretation of the curriculum. This led to the use of marking and scoring procedures that aligned with what teachers might do in classroom assessments. As the goals of the assessments for Grades 3 and 6 have shifted, especially with the added requirement for target setting by schools, EQAO has adjusted its marking and scoring procedures. The Grade 9 assessment and the OSSLT have been successfully developed for their particular requirements. At this time, further consolidation of the procedures across the assessments is in order. This is especially important to gain efficiency in the use of EQAO technical staff on the one hand and in communicating results and procedures on the other. Also, the state of the psychometric art has improved since the early years of EQAO, and improvements in accuracy and efficiency are now possible.

Finding 6. The EQAO assessment programs now use a variety of approaches to marking and scaling students’ work. These approaches, while adequate, may not be optimal for producing the best quality results from the students’ work.

Recommendation 6. That EQAO develop a unified system for marking and scaling students’ work. That the marking be based on item-specific marking guides that align with the nature of the anticipated responses, not the final reporting scale. That the statistical scaling use item response theory models – preferably the Rasch family of models. That alignment of the scale to the reporting scale be based on formal standard setting procedures. That EQAO rigorously document all of its procedures for marking and scaling the students’ work for each assessment program.

EQAO is asked to produce many kinds and levels of reports, from individual student reports that go to students and their parents, to school reports used for school improvement planning and target setting, to board and provincial reports used for accountability and curriculum evaluation. All large-scale assessment programs run into conflicting goals concerning the tradeoffs of detail and accuracy in reporting; this is a major unsolved problem in the field of applied measurement. In the first years of the Grade 3 and 6 assessments, EQAO was obliged to present details at the level of categories and strands for students, schools, boards, and the province. But it is now clear that the test structure (especially, the number of items per category) cannot support such detail within the current test designs. As a result, EQAO has been cutting back the reports to a level of detail that the tests can accurately provide.
Finding 7. The EQAO assessment programs have sought to produce detailed reports for students, schools, and boards. However, some of the results have been based on very few items. Such results are neither as reliable nor as representative of the strand or category as they should be to be truly useful to students, teachers, or school administrators.

Recommendation 7. That EQAO define specific requirements concerning comprehensiveness and detail of the measurement and concerning accuracy and comparability of scores and reports at the individual, school, board and provincial levels. That EQAO consider alternative approaches, especially item mapping, to reporting information and providing interpretation of scales.

The first years of EQAO were focused on getting complex, large-scale assessments designed and implemented. The assessment designs were motivated by short- and long-term educational program requirements: improving evaluation practices and providing better accountability in Grades 3 and 6, assuring and tracking the implementation of the new Grade 9 mathematics curriculum, and establishing a literacy standard for high-school graduation and encouraging higher literacy achievement. EQAO has succeeded in getting each assessment established. That is an exceptional accomplishment: Many educational assessment systems in other jurisdictions have failed to get past an initial survey. Continued success at EQAO will depend on adequate long-term planning.

Finding 8. The EQAO assessment programs have many strengths. By following the recommendations in this report, they can be improved. However, future reviews will be needed, and an expanded, on-going program of research and development should be planned. The long-term goals of EQAO should be examined.

Recommendation 8. That EQAO strengthen the ways it carries out research, obtains advice, and plans for the long term:

• That the assessment program be supported by an active program of validity research.
• That there be improved long-term technical planning for the EQAO assessment programs and systems.
• That EQAO have a regular program of obtaining reviews and advice. That there be a high-level technical advisory panel with a regular schedule of meetings.
• That EQAO determine the time frame for its assessment programs and for periodic reviews. That EQAO carry out a major review every 3 to 5 years.
• That EQAO adopt assessment designs that are sufficiently general that they provide useful bridges across curriculum changes.
• That EQAO collect critical information to provide longer-term trend data.
In summary, we believe that EQAO has created world-class educational assessment programs. Our analysis and critique is aimed at identifying the strengths of the assessments and pointing out areas where clarification of purposes and constructs, improvement of systems and methodologies, and incorporation of new analysis and reporting technology could enhance EQAO's work, which has become a vital part of Ontario education.

Richard Wolfe
Ruth Childs
Susan Elgie

Toronto, 26 April 2004
CONTRIBUTORS

Principal Investigators
Richard Wolfe, Ontario Institute for Studies in Education of the University of Toronto
Ruth Childs, Ontario Institute for Studies in Education of the University of Toronto
Susan Elgie, Ontario Institute for Studies in Education of the University of Toronto

Liaison with EQAO’s Chief Executive Officer
Kenneth Leithwood, Ontario Institute for Studies in Education of the University of Toronto

Technical Advisors
Alister Cumming, Ontario Institute for Studies in Education of the University of Toronto
James P. Cummins, Ontario Institute for Studies in Education of the University of Toronto
Karen L. Draney, University of California – Berkeley
Lorna M. Earl, Ontario Institute for Studies in Education of the University of Toronto
Ronald K. Hambleton, University of Massachusetts
Don A. Klinger, Queen’s University
Dany Laveault, University of Ottawa
Alexandra Lawson, Lakehead University
Cinde L. Lock, Ottawa-Carleton District School Board
Pamela A. Moss, University of Michigan
David R. Olson, Ontario Institute for Studies in Education of the University of Toronto
Shelley Peterson, Ontario Institute for Studies in Education of the University of Toronto
Barbara S. Plake, University of Nebraska – Lincoln, Buros Institute of Mental Measurements
Mark Reckase, Michigan State University
William H. Schmidt, *Michigan State University*

Christine A. Suurtamm, *University of Ottawa*

Merrill Swain, *Ontario Institute for Studies in Education of the University of Toronto*

Ross E. Traub, *Ontario Institute for Studies in Education of the University of Toronto*

Mark R. Wilson, *University of California – Berkeley*

Robert J. Wilson, *Queen's University*

Laurell L. Wise, *Human Resources Research Organization, Alexandria, VA*

**Researchers**

At the Ontario Institute for Studies in Education of the University of Toronto:

  Zoi Angelis, Sonia Ben Jaafar, Marcella Campbell, Jennifer Dunn, Barnabas Emenogu, Olesya Falenchuk, Tahany Gadalla, Monique Herbert, Vandhana Luthra, Anjali Mazumder, Jodi Herold McIlroy, Yunmei Xu, Wendy Yen, Su Zhang

At the University of Ottawa:

  Patrick Moisan

At Queen's University:

  Tess Miller
ACKNOWLEDGEMENTS

Many people contributed to the EQAO Assessment Review Project. The technical advisors and researchers are listed on the preceding pages. They spent countless hours reviewing materials, analysing data, and discussing possible designs for the assessment programs.

We are grateful to EQAO staff for providing extensive information, advice, and ideas throughout the project.

Conducting a large project involves policies and paperwork. We were fortunate to have the assistance of Lara Cartmale and Bessie Giannikos in OISE/UT’s Department of Curriculum, Teaching and Learning, as well as the help of Amanda Barrett, Lynette Davis, Pampa Dey, and Ava Lew. Finally, for editorial assistance, thank you to Maureen Murney, and for help in assembling this final report, to Nina Lewis, Michelle Pon, and Melissa White.

Richard Wolfe, Ruth Childs, & Susan Elgie
CHAPTER 1
INTRODUCTION

ONTARIO’S ASSESSMENT PROGRAMS
In 1996, the Ontario government established the Education Quality and Accountability Office (EQAO) to “…develop tests and require or undertake the administering and marking of tests of pupils in elementary and secondary schools … [and] report to the public and to the Minister of Education and Training on the results of tests and generally on the quality and effectiveness of elementary and secondary school education and on the public accountability of boards” (Education Quality and Accountability Office Act, 1996, S.O. 1996, c. 11).

EQAO now has four assessment programs, developed and administered in both French and English:

- Grade 3 Assessment of Reading, Writing and Mathematics (since April 1997)
- Grade 6 Assessment of Reading, Writing and Mathematics (since May 1999)
- Grade 9 Assessment of Mathematics (since January 2001)
- Ontario Secondary School Literacy Test (OSSLT; since October 2000)

These assessment programs are administered in every publicly-funded elementary and secondary school in Ontario. About a third of Ontario’s school-aged children participate each year. Consequently, a third of Ontario’s teachers are involved in administering the assessments. Every family in Ontario with school-aged children receives results from an EQAO assessment at least once every three years.

This is one of the most extensive systems of performance-based educational evaluation in the world. Extraordinary efforts have been made to connect the tests to the curriculum contents, to maintain authentic links to classroom practices, and to attend to the multiplicity of purposes of assessment, including individual reporting, modelling evaluation practice, encouraging capacity building around the curriculum and evaluation, and providing information for school improvement and school and board accountability.

In the seven years since EQAO created its first assessment program, new curriculum documents have been introduced (and are now being reviewed), along with new report cards. A requirement that schools set targets for “the percentage of Grade 3 students who will meet the provincial standard for reading in the EQAO reading assessment” has been added. For the 2003-04 school year, the Ministry of Education created a new course, the Ontario Secondary School Literacy Course. The most effective assessment programs are designed to fit specific contexts. When the context changes, the wise developer revisits the design. This report describes such a revisiting – an independent review of the assessment programs – commissioned by EQAO.

1. In addition, EQAO has responsibility for other assessment surveys, indicator programs, and participation in national and international studies. This work was not part of our review.
THIS REVIEW

As explained in the Foreword, EQAO embarked on a strategic initiative, “Ensuring Quality Assessments,” in November 2002. This was envisioned as a systematic and cyclical process for updating EQAO practices and aimed at ensuring that the assessment processes incorporate recent improvements in standards and practices in large-scale assessment and that reporting meets user needs for demonstrating accountability and improvement planning. The initiative comprised research and study, consultation, analysis and synthesis and implementation. EQAO carried out internal review, study of other jurisdictions, and consultation with stakeholders.

In March 2003, EQAO requested quotations for an external project to review the assessment programs, in particular the testing and reporting practices. The Request for Quotations (RFQ-002-005) specified the purpose was to “provide EQAO with rigorous bases and systems for implementing new test and report designs and for guiding practices over the next five years.” EQAO subsequently expanded this description to: “In consideration of current standards for large-scale assessment, review, refine and/or affirm all processes associated with EQAO assessment programs for Grades 3, 6, and 9 and the OSSLT.”

In April 2003, EQAO awarded a contract (C-03-003) for the review to the Ontario Institute for Studies in Education of the University of Toronto (OISE/UT).

The review was led by OISE/UT professors Richard Wolfe and Ruth Childs and by the head of OISE/UT’s Research Consulting Service, Susan Elgie. A team of internationally renowned experts in educational assessment was assembled to contribute to the review.

The review commenced in June 2003. In a series of meetings with EQAO, we developed an interactive method of continuing specification of the purposes and activities of the project. At each meeting, but particularly after the large consultative meetings and after delivery of draft reports, we engaged in intense exploration of the many topics under the purview of the review. This interactive method facilitated an evolving and adaptive approach to the task which made our work much easier and we hope also increased its relevance and usefulness.

To facilitate the review, EQAO provided access to its assessment materials, documentation, and staff. EQAO also arranged discussions with the educators taking part in its marking sessions. We had access to electronic files of assessment data, so that we could conduct independent analyses of students’ scores.

STANDARDS FOR ASSESSMENTS

An assessment program is much more than the items that students across Ontario answer when they sit for the assessment. It includes how the items are developed and the assessment forms are assembled, the instructions for administering the assessments, the process for marking the students’ work, the scaling of the marks, and how the results are reported. Over the ten months of the project, we reviewed all these aspects of the assessment programs.
What does an excellent assessment program look like? In 1993, the *Principles for Fair Student Assessment Practices for Education in Canada* was developed by representatives from the Canadian Education Association, the Canadian School Boards Association, the Canadian Association for School Administrators, the Canadian Teachers’ Federation, the Canadian Guidance and Counselling Association, the Canadian Association of School Psychologists, the Canadian Council for Exceptional Children, the Canadian Psychological Association, and the Canadian Society for the Study of Education. In addition to general guidelines for student assessments, the *Principles* includes a section outlining six principles for mandated assessment programs:

Developers and Users should:

1. Inform all persons with a stake in the assessment (administrators, teachers, students, parents/guardians) of the purpose(s) of the assessment, the uses to be made of the results, and who has access to them.

2. Design and describe procedures for developing or choosing the methods of assessment, selecting students where sampling is used, administering the assessment materials, and scoring and summarising student responses.

3. Interpret results in light of factors that might influence them. Important factors to consider include characteristics of the students, opportunity to learn, and comprehensiveness and representativeness of the assessment method in terms of the learning outcomes to be reported on.

4. Specify procedures for reporting, storing, controlling access to, and destroying results.

5. Ensure reports and explanations of results are consistent with the purpose(s) of the assessment, the intended uses of the results and the planned access to the results.

6. Provide reports and explanations of results that can be readily understood by the intended audience(s). If necessary, employ multiple reports designed for different audiences. (Joint Advisory Committee, 1993, p. 20)

We will refer to these principles throughout this review.

**THIS REPORT**

This report has eight chapters. Chapters 2 through 7 focus successively on the stages of development of an assessment program: clarifying purposes, defining constructs, designing tests, developing items, scoring, and reporting. In each chapter we begin by briefly stating the context, relating the EQAO work to best practices in assessment, and mentioning the strengths and weaknesses of the EQAO work. Then we summarise one or more findings and recommendations. The balance of the chapter is devoted to elaborating the context of and reasons for making the recommendations, based on the results of the review. Chapter 8 concerns recommendations for ensuring continued improvement and success.
There are four appendixes to this report. Appendix A contains a technical history of equating in the Grade 3 and assessment. Appendix B illustrates the item response theory scaling that is proposed as a general approach for EQAO assessments. Appendix C illustrates how matrix sampling can enhance information in an assessment. Appendix D contains a glossary of some specialised terms that are used in assessment design.

At various points, we describe steps EQAO is already taking to ensure the quality of its assessment programs. During the review, the work of EQAO has gone on – new assessment materials have been developed, assessments have been administered, marked, and scaled, and results have been reported. Throughout the review, we have provided preliminary advice on many of the operational activities and EQAO has made some important refinements. Therefore, in addition to recommending future changes, we will describe the changes that EQAO has begun to make, in response to this review and to its own internal review, survey of practices in other jurisdictions, and consultation with stakeholders.

We believe that assessment programs such as those developed and administered by EQAO can provide important information for students, parents, and educators. We also believe that any assessment program can be improved. The recommendations we offer in this report are based on ten months of talking with EQAO’s staff, examining EQAO’s data and documentation, and seeking advice from other experts. It is our belief that these recommendations will help Ontario continue to have the best assessment programs possible.
CHAPTER 2
CLARIFYING THE PURPOSES OF THE ASSESSMENT PROGRAMS

Classroom teachers assess students’ achievement every day. The assessments gather information about what students know and can do. In addition, some of the assessments may help teachers identify students’ misunderstandings or plan instruction. Some will help teachers assign final marks. Some assessments will help students better understand their own learning processes. Teachers do not try to accomplish all of these purposes with every assessment. They also do not try to assess every part of the curriculum with a single assessment tool. Instead, teachers design assessments with specific curriculum expectations and specific purposes in mind.

Large-scale assessments, such as those developed by EQAO, also gather information about what students know and can do. Such assessments may have other purposes as well, such as to provide models for classroom assessments or to help schools plan instructional activities and set student achievement targets. Like a classroom assessment, a large-scale assessment must be designed with a particular purpose in mind. For example, modelling classroom assessment may require a different assessment design than monitoring achievement of targets.

As the purposes of educational assessments expand, test design might be compromised between radically different and incompatible requirements. The best practices for one purpose might be inconsistent with those for another. For example, information returned to students needs to have highly reliable scores; information useful for teachers or for curriculum planning needs to have detailed content articulation; and information useful for monitoring and accountability needs to have stable measurement over time. Reliability, detail, and stability are all positive attributes, but one test design cannot excel at all three. In some jurisdictions, separate testing programs are used for distinct purposes, although a current methodological challenge in the field of educational measurement is to develop testing technologies that serve multiple purposes well.

Our analyses of the historical and current purposes of the EQAO assessments, as described by EQAO and evident in its reporting and interpretative material, show clear but shifting and expanding intentions and, we are afraid, situations where the accepted purposes have grown too extensive or diffuse. In this chapter, we compare those intentions with the technical characteristics and capabilities of the test designs.

Finding 1. The assessment programs have been asked to serve a myriad of purposes. EQAO has tried to accommodate these purposes. However, because it is impossible to design an assessment program to meet diverse purposes, the proliferation of purposes may compromise how well EQAO’s assessments can accomplish any one of them.

Recommendation 1. That EQAO continue to develop, in consultation with stakeholders, clear statements of the purposes of each assessment program, including the relative importance of each purpose.
EQAO’S MANDATE

As we mentioned in Chapter 1, the bill that legislated testing in Ontario specified that EQAO should “…develop tests and require or undertake the administering and marking of tests of pupils in elementary and secondary schools … [and] report to the public and to the Minister of Education and Training on the results of tests and generally on the quality and effectiveness of elementary and secondary school education and on the public accountability of boards” (Education Quality and Accountability Office Act, 1996, S.O. 1996, c. 11). The EQAO Act did not specify what grades or subjects should be tested. In announcing the Act, however, then Ontario Minister of Education, John Snobelen, filled in some of the details: “In the 1996-97 school year, [EQAO] will conduct the first annual test of all Grade 3 students in reading, writing and mathematics, along with a sampling of students in Grade 6 or 9, … testing will give us valuable, accurate and credible information on how students are performing in schools” (Ministry of Education News Release, December 14, 1995). As we described in Chapter 1, assessments of all students in Grade 6 in reading, writing, and mathematics, of all students in Grade 9 in mathematics, and of high school students, beginning in Grade 10, in literacy, have since been added.

Defining the purpose of an assessment program must go beyond identifying the students to be tested and in which subject areas. The wording in the original mandate provided some hints. EQAO was to report:

- on the results of the tests,
- on the quality and effectiveness of elementary and secondary school education, and
- on the public accountability of boards.

The first of these, reporting on the results of the tests, seems straightforward: Students’ performance on the tests should be reported to the public and the Minister. However, as we will discuss later in this report, there are many, many different ways in which test results can be reported. Should the results for each item be reported or summarised across groups of items? If items are to be grouped, how? How much detail should be reported for individual students, for schools and boards? To make these decisions, we need to know how the results will be used. That is, we need to know each assessment program’s purposes.

The second requirement, reporting on the quality and effectiveness of elementary and secondary school education, provides some insight. One purpose, it seems, is to judge the “quality and effectiveness” of Ontario’s schools. The assumption is that, if schools provide a good quality of education, then students will perform well on the assessments. Therefore, the reasoning goes, if students perform well on the assessments, then schools must be providing a good education.

The third requirement, reporting on the public accountability of boards, suggests an emphasis on the results produced by boards in relation to the resources they spend. This purpose requires that results be reported for boards, but not necessarily for individual students.
Since the original mandate, the Ministry of Education has added other purposes. Schools are required to set reading assessment targets for their Grade 3 students, and so EQAO must produce results that can be used in this way. The OSSLT is a graduation requirement beginning with the class that will graduate in June 2004, and so EQAO must produce results that are appropriate for use with such high stakes.

EQAO also tries to achieve other purposes. For example, a goal of the assessments in Grades 3 and 6 has been to help teachers engage with the new curriculum and to model good classroom practice. EQAO has used the summer marking sessions as opportunities to help teachers improve their assessment skills and learn about the EQAO assessments.

In fact, EQAO now lists three purposes for its assessments:

- Accountability
- School/System Improvement
- Capacity Building

What does EQAO mean by these purposes? Under the first purpose, accountability, EQAO lists questions such as: Are students learning the curriculum? Are students reaching the provincial standards? Are results improving over time? Under the second, school/system improvement, EQAO seeks to answer: What can schools do to improve student results? By capacity building, EQAO means: Do the assessments support teachers’ professional development?

**PURPOSES AND TEST DESIGNS**

Professor Robert Wilson of Queen’s University wrote in a paper commissioned for this review:

> It is axiomatic in the educational testing and assessment world that individual tests may accomplish one, or perhaps, two purposes well, but that attempts to extend their functions beyond that will almost certainly hamper their utility. (Wilson, 2003, p. 1)

To understand why this is so, one need only think about the many decisions that must be made in designing any assessment program:

- What constructs will the assessments measure and how are they connected to the curriculum?
- When will the assessment be administered?
- What types of items will it have? How many items of each type?
- What content will the items measure?
- How long will the assessment take to administer?
- Who will mark the responses?
• How will the students’ responses be marked?
• How will the marks be translated to results to be reported?
• How will the results be made comparable across administrations?
• How many and which items will be released?
• How will the results be reported? How much detail will be provided at the student, school, board, and provincial levels?

To each of these and the many other questions that must be answered in designing an assessment program, several answers are possible. In order to make any of these decisions, a test developer must consider the purpose of the assessment program. An assessment program that is very good for providing detailed student results will not necessarily provide results that are useful for comparing school results across years. One that is a high school graduation requirement and so must yield a pass-fail decision will need to be designed quite differently from an assessment program that is intended to distinguish more levels of student performance. Providing detailed diagnostic information will require a longer assessment than simply providing a pass-fail decision.

The purpose of an assessment program drives its design. Lack of clarity about an assessment’s purpose will make it difficult to determine an appropriate design. The importance of defining the purpose is recognised in the Principles for Fair Student Assessment Practices for Education in Canada: Assessment programs should “inform all persons with a stake in the assessment (administrators, teachers, students, parents/guardians) of the purpose(s) of the assessment, the uses to be made of the results, and who has access to them” (Joint Advisory Committee, 1993, p. 20). In addition, any design is unlikely to be optimal for multiple purposes. Designing an assessment program to meet multiple purposes invariably requires difficult compromises.

FINDINGS
The Grade 3 and Grade 6 assessments have seen the greatest expansion of purposes. Their initial designs were intended to model excellence in classroom assessment. Over the seven years, there has been a “mandate spread.” Because there is an important lesson here in the interaction of purposes and design, we provide a detailed history in Appendix A. Here is a summary:

1. The assessment for each grade is related to a theme, “change” in Grade 3 and “perspectives” in Grade 6. Within these broad themes are specific themes each year. The in-class work is organised with a student magazine and work booklets for reading, writing, and mathematics. All materials are released each year.
2. The limitation of this design is that it makes compatible measurement between years very difficult, since no part of the assessment can be repeated, both because of the release of material and because of the incompatibility of theme from one year to the next. Yet it is inevitable that the different items and the test as a whole will be somewhat harder or somewhat easier. To measure real improvement or decline between years, we have to be able to distinguish how much of the change is due to the particular items and test and how much is due to real changes in students' performances. Perhaps in the early years of the assessment, this was not so important, but with the advent of target setting and public rankings, it is essential.

3. To equate the tests, EQAO must have either (a) some of the same items in both years or (b) some of the same students. Neither is feasible in the base design. To provide items that could be repeated across years, another section was added to the Grade 3 and 6 assessments: a booklet of multiple-choice items that were not related to the theme. There have been four versions of the multiple-choice booklet each year. Each student received one of the four versions. Each year, three versions were kept secure and one was retired. The retired booklet was then replaced with a new booklet the following year.

4. Since three-quarters of the multiple-choice items are the same from year to year, the students' scores on the multiple-choice can be statistically equated and “true” change measured. Of course, this is change in that aspect of reading that is measured by the multiple-choice test. It overlaps with what is measured by the performance part of the test and is a part of the overall construct that EQAO is trying to measure, but it is incomplete. The crux of the problem is that EQAO had no choice but to base the equating for the whole EQAO Grade 3 and 6 assessments on the equating of the multiple-choice items. The alternative was to report raw, unequated performance scores, and that would have introduced uncontrolled, random, and doubtless wrongly interpreted fluctuations from year to year.

The above example illustrates the difficulty of meeting competing purposes. The purpose of modelling classroom assessment resulted in a theme-based assessment and a decision that students' booklets should be returned to the schools each year. What we hear anecdotally is that some teachers have used the Grade 3 and 6 assessments as models for their own classroom assessments. The purpose of measuring change between years was not considered a priority in the original design. That purpose became more important especially because of the added-on use of the assessment in the target setting as mandated by the Ministry of Education, but also because of public attention to school-by-school rankings and changes. But the theme-based design had already been established and it limited the possibilities for making the results comparable across years. The link formed by the multiple-choice items is a compromise solution, probably the best available given the diverse purposes.

The Grade 9 assessment and the OSSLT must also try to meet incompatible purposes. For example, the most important purpose of the Grade 9 assessment seems to us to be the production of results to help schools plan instruction. Producing detailed results for each individual student has limited the design options for the Grade 9 assessment, eliminating
some designs that could produce results of more use to schools. Similarly, the OSSLT serves as a graduation requirement, so must distinguish between students who pass and those who need remediation. As we will discuss later in this report, for those students who fail it, the OSSLT has also provided feedback to schools about individual students’ areas of strength and weakness. We question whether this feedback is either psychometrically or practically useful.

EQAO has recently begun moving toward a design for the Grade 3 and 6 assessments in 2005 that puts less emphasis on modelling classroom assessment and more emphasis on measuring change over time. We believe that such a change would be facilitated by a clear statement of the purpose of the assessment – and, if there are to be multiple purposes – of the relative importance of each.

**CONCLUSION**

Because any assessment program must be designed to fit its purpose, defining the purpose is of utmost importance. The original mandate for EQAO included several purposes, and more have since been added. In addition, the purposes have not been clearly stated or prioritised. No assessment program can simultaneously accomplish many purposes well; even less so, when the purposes are not clearly defined. Therefore, we recommend that EQAO continue to work, in consultation with stakeholders, to clearly define, prioritise, publish and publicise the purposes of EQAO’s various assessment programs. We suspect that a reduction in the number goals will be necessary.
CHAPTER 3
DEFINING THE CONSTRUCTS AND SPECIFYING THE CURRICULUM CONNECTIONS

When a teacher creates an assessment, she goes through a series of steps. She decides, for example, that a written exam will help her evaluate whether students have mastered the concepts and skills covered in the latest unit. She then considers the importance of the exam (the proportion it will represent of an overall grade), selects the content to be measured and the types of questions to include, decides how long the students will have to answer the questions, how she will mark it, and how to communicate the results to the students. She makes these decisions based on her knowledge of the curriculum and of her students.

Many of these decisions are also required in large-scale assessment programs. However, while the teacher need not formally rationalise and document her decisions, a large-scale assessment program must. Every decision, ranging from the curriculum expectations included on the test to how the students’ work will be marked and scaled, must be described in detail. This is important because, unlike classroom assessment where the teacher independently conducts and marks the test, large-scale assessments involve many individuals, all of whom must implement consistent decisions. Furthermore, these decisions must be available for scrutiny and discussion. In this chapter, we describe the first important decisions that must be documented by any assessment program: the construct or constructs to be measured and the connections to the curriculum.

Assessment programs here and in other jurisdictions strive in best practices to base tests on official, mandated curriculum. But here as elsewhere this is not straightforward, because the curriculum specifications generally deal with content and instructional specifications and not with evaluation. They do not provide a sufficiently detailed specification of the constructs that are to be measured in the assessments. There is need for intermediate analysis and documentation to link constructs and test specifications with the curriculum. In this chapter, we discuss the processes by which EQAO produces test frameworks and specifications and the results of those processes.

Finding 2. EQAO has created framework documents based on the Ministry curriculum. However, these framework documents neither clearly define the construct or constructs each assessment is intended to measure nor describe with sufficient detail the decisions that were made in interpreting the curriculum documents and developing the assessment frameworks.

Recommendation 2. That EQAO review, revise, and expand documents that define the constructs to be measured by the assessments and the connections between the assessments and the curriculum.
CONSTRUCT DEFINITIONS

The first issue an assessment program must address and describe is the construct or constructs it is attempting to measure. “Construct” is a psychological term used to refer to something theorised to exist. Mathematical understanding and literacy are two examples of constructs. Constructs cannot be observed directly; rather, they are inferred through observation of behaviours (e.g., responses to test items). Therefore, before we can hope to measure them, we must be clear about what it is we are measuring.

Defining a construct is a difficult, complex, and often, controversial task. Take the OSSLT, for example. According to the Highlights of Provincial Results, April 2003, “The OSSLT is designed to assess the reading and writing skills students are expected to have learned across all subjects by the end of Grade 9, as outlined in The Ontario Curriculum.” The definition of literacy used in creating the OSSLT seems to be mastery of the reading and writing skills required in the Ontario Curriculum.

However, this definition does not clearly correspond to previously outlined definitions of literacy. The 1995 report of the Royal Commission on Learning, For the Love of Learning, recommended that Ontario create an assessment to “assure the public that a high school diploma signals adult literacy; that no high school graduate is incapable of reading and writing well enough to communicate in a post-secondary classroom, on the job, or in order to meet the demands of everyday life as a citizen and voter” (vol. 2, p. 151). Professor Robert Wilson, in a paper for this project, characterises this latter definition as “functional literacy.” Wilson notes that there is some evidence of a concern for functional literacy in the OSSLT:

Some support for the idea that this test concentrates on functional literacy can be gleaned from its specifications. The writing prompts, for example, are utilitarian: a summary, an opinion piece, an informational paragraph, and a news report. Similarly, the reading passages are composed of a narrative, an informative piece, and a graphical display, all deemed practical in terms of extra-school activity. (Wilson, 2003, p. 1)

Defining the constructs of reading, writing and mathematics to be measured by the Grade 3, 6, and 9 assessments is equally important. As with the OSSLT, there is a temptation to define these constructs based entirely on the curriculum documents. While clear connections with the curriculum are very important, curriculum documents may not provide sufficiently clear construct definitions that can be used as the foundation of an assessment’s design. Further, as several of the technical advisors to the review pointed out, curriculum documents are periodically revised. For example, in a paper commissioned by this project, Dr. Cinde Lock, of the Ottawa-Carleton District School Board, and Professor Robert Wilson of Queen’s University recommended that EQAO develop for the Grade 3 and 6 assessments:

[A] conceptual reporting structure that is based on critical aspects of Reading and Writing development in children and not initially derived from the curriculum itself.
In this model, the curriculum will still be aligned with the assessment, but decisions about which aspects of the curriculum will be assessed and reported would rely more heavily from a determination of what is deemed important to the discipline. Following this model, the assessment program may need to be revised less often, as curriculum documents periodically are changed, and the capacity of the program to measure improvements across the system would be enhanced. (Lock & Wilson, 2003, p. 17)

As Dr. Lock and Professor Wilson point out, however the constructs are defined, it is important that the assessments’ connections to the curriculum are specified. The establishment of these connections will be discussed next.

**CURRICULUM CONNECTIONS**

Once the construct has been defined, an assessment framework must be developed. The creation of an assessment framework involves deciding what specific content should be assessed. For example, EQAO’s *OSSLT Curriculum Connections* document defines some of the connections between the assessment’s content and the curriculum. However, as noted by Professor Wilson, there is a tension between the definition of literacy as reading and writing in the curriculum and its definition as practical skills needed after graduation. In fact, he suggests that students’ performances on the assessment might be seen as a critique of the curriculum:

> To the extent that students are or are not able to do these things might also be a fair comment on the degree to which the curriculum provided to teachers and students, however well implemented, produces individuals who have developed these skills. In short, the test results would be an indicator whether or not the curriculum is adequate to one critical aspect of its overall purpose. (Wilson, 2003, p. 2)

The Grade 3 and Grade 6 Assessments of Reading, Writing and Mathematics are based on two of the three strands from the language curriculum – Reading and Writing (the third strand is Oral and Visual Communication) – and on the five strands defined in the mathematics curriculum – Number Sense, Geometry, Measurement, Pattern/Algebra and Data/Probability. In addition to strands, the curriculum also defines categories, which relate to the skills students need to develop. In reading and writing, the categories are: Reasoning, Communications, Organization of Ideas and Applications of Language Conventions. In mathematics, the categories are: Problem Solving, Understanding Concepts, Application of Procedures and Communications.

The Grade 9 Assessment of Mathematics is based on *Ontario Curriculum Grades 9 and 10, Mathematics*, which describes two courses at Grade 9: Principles of Mathematics (Academic) and Foundations of Mathematics (Applied). The courses are described as different in approach, but similar in content. In the Grade 9 curriculum, the “expectations” statements regarding what the student would know and be able to do at the end of each course are organised in four strands: Number Sense and Algebra, Relationships, Analytic Geometry,
and Measurement and Geometry. Across these four strands, teachers in both courses were expected to assess four categories of knowledge and skill: Knowledge/Understanding, Thinking/Inquiry/Problem Solving, Communication, and Application.

**Framework Documents**
Although selecting strands and categories from the curriculum documents is quite easy, producing a more detailed framework is very difficult. The curriculum documents are intended to guide classroom instruction, not to serve as frameworks for large-scale assessments. Interpreting these documents for use in creating large-scale assessments requires many decisions.

For example, in their papers, Dr. Lock and Professor Wilson outlined the results of decisions that were made when creating the reading and writing parts of the Grade 3 and 6 assessments:

- Overall and specific Ministry expectations have been mixed together and not prioritised in the corresponding listing of EQAO expectations.
- Ministry expectations have been omitted from the EQAO framework. Often these omissions are justified because such expectations would be difficult, if not impossible, to assess in a large-scale setting. Other omissions have been made for unknown reasons. For instance, it is unclear why a Ministry expectation such as: “Begin to develop their own opinions by considering some ideas from various written materials” or “Accurately use appropriate organisers (e.g., table of contents, index)” have not been included anywhere in the EQAO framework for the assessment.
- Additional curriculum expectations have been created for the EQAO framework that do not appear in the Ministry curriculum, such as: “Produce writing that reflects their capacity for independent, critical thought,” and “Make inferences.”
- Ministry expectations have been subdivided and included as multiple expectations in the EQAO framework. For instance, “Summarize and explain the main ideas in information materials and cite details that support the main ideas” has been split into two EQAO expectations.
- Multiple Ministry expectations have been grouped together and included as single expectations in the EQAO framework. For instance, in the Grade 3 Writing strand 15, specific expectations such as “Use common prefixes and suffixes” and “Use irregular plurals correctly” have been subsumed under one overall expectation: “Use correctly the conventions specified for this grade.”
- Ministry expectations have been edited, made more concise and in some instances, restated in such a way that the original intent of the expectation has been altered. For instance, the Ministry expectation, “Uses their knowledge of the characteristics of different forms of writing to select the appropriate materials for a specific purpose” has been changed to “Use knowledge of organization and characteristics of forms of writing as a guide to reading” in the EQAO framework.
The difficulty of interpreting the curriculum documents and the framework documents is further illustrated by the work of Professor Christine Suurtamm of the University of Ottawa. In a study commissioned for this review, Professor Suurtamm and two colleagues independently matched items from 2003 forms of the Grade 9 assessment to the curriculum expectations and categories. She reported that the expectations to which her team and EQAO staff assigned the items agreed for only 60% of the items and the classifications by category agreed for 63% of the items. Further, she found the following:

- The greatest discrepancy was in comparing the raters’ and EQAO’s coding of multiple-choice questions. This is true for both the categories of the Achievement Chart and for the curriculum expectation match.
- Furthermore, the multiple-choice questions contained several questions that did not match the Grade 9 curriculum. For instance, one question made use of properties of similar triangles, which is not introduced until the Grade 10 mathematics curriculum.
- Even though items were matched to expectations, some of the items did not address the full expectation. An item is often matched to a curriculum expectation that addresses the “content” of the curriculum expectation but not the suggested mathematical “operation.” (Suurtamm, Moisan, & Luthra, 2004b, pp. 3-5)

Professor Suurtamm concluded:

Our results show that there is concern over whether the Achievement Chart categories can be clearly defined and separated to a degree suitable for reporting on each category. Matching to individual curriculum expectations is also problematic. Items are more easily matched to a small cluster of Specific Expectations or an Overall Expectation. (Suurtamm, Moisan, & Luthra, 2004b, p. 5)

Most likely, each of the decisions that resulted in the assessment framework was carefully considered. The difficulty, however, is that the decisions are not described in the framework documents – indeed, Dr. Lock and Professor Wilson detected the differences between the curricula and the frameworks only by a careful comparison of the documents.

Furthermore, these decisions are not limited to the curriculum expectations. Each curriculum document provides an “Achievement Chart” describing four levels of performance for each of the categories in that subject. Based on the Achievement Charts, EQAO has developed “Assessment Scales,” in which descriptions of student performance are limited to the types of evidence that EQAO collects (in particular, they do not refer to performance demonstrations over time or in groups). Of course, the decisions made in creating the Assessment Scales also must be documented.
**Bridge Documents**

It may be possible to document all decisions and their rationales in the framework documents. However, EQAO should consider creating additional “bridge documents.” Some large-scale test developers (e.g., the state of Massachusetts for its Massachusetts Comprehensive Assessment System) create such documents to link the curriculum documents and the framework documents. In a bridge document, the interpretation of the curriculum documents is described in detail, with justifications for each decision in that interpretation. One advantage of creating, for each assessment program, a bridge document that is separate from the framework document is that the bridge document can focus on the decisions and rationales and leave the details of how many items of each type will be included on the assessment to the framework document. Bridge documents are sometimes circulated to other educators for comments. The best bridge documents are the result of an intensive process of consultation and consensus building. The creation of such a document can provide opportunities for stakeholders to discuss key interpretations and decisions.

**Blueprints**

A part of a framework document that merits additional discussion is the test blueprint. The blueprint, as the name suggests, provides sufficient specificity that it is possible to actually construct a version of the assessment based on the document. Where the rest of the framework document – and the bridge document – record the decisions and their reasons, the blueprint details how many items of what types are needed and which parts of the curriculum they are to measure. The blueprint should also include instructions for balancing curriculum coverage across items and specify which items are to be released and which are to be used to link results across years. If different items are to be administered to different students, it should specify which items are to be rotated with what sampling density.

**FINDINGS**

In addition to the results mentioned earlier, a number of issues surrounding the current construct definitions and curriculum specifications were identified. Professor Alexandra Lawson of Lakehead University examined the Grade 3 and 6 mathematics items and expressed concern about how the curriculum documents were being used as the basis for an assessment. In a report for this review, she wrote:

> The 1997 curriculum… outlines a list of expectations and some general instructional methods and emphases but offers limited direction on what is important (the ‘Big Ideas’), how these ideas connect and some continuum or progression of development. The lack of a larger picture and some prioritising of concepts makes it difficult to determine whether or not the important ideas are being assessed. (Lawson, 2003, p. 8)

Professor Lawson recommended that EQAO make deliberate decisions about which parts of the curriculum should receive more attention in the assessments, rather than treating all expectations as equally important. This was echoed by teachers who were involved in marking the assessment. As one teacher, in a focus group for this review, put it:
I find different strands tend to be weighted differently in the classrooms as well. The number sense and numeration, it takes up the bulk of the classroom time, whereas patterning and algebra have very few overall specific objectives and are given a lot less emphasis in the classroom. There’s a couple of strands that are given much less emphasis. (Elgie, 2004, p. 4)

This issue is also reflected in the OSSLT, by the question of whether reading and writing should be treated as distinct constructs. In fact, the OSSLT is inconsistent in this regard: Reading and writing are measured within a single test, but separate scores are produced. This inconsistency has caused practical problems in test administration. In particular, students who receive only one passing score must sit again for the entire test, which suggests a possible test-taking strategy of focusing on only the reading items on one administration and on the writing items on the next. A review of jurisdictions with literacy tests found that some had a single literacy test yielding a single score and some had separate reading and writing tests; none but Ontario had one test and two scores.

Professor Shelley Peterson, in an examination of the OSSLT for this review, noted that the OSSLT’s emphasis on reading and writing across the curriculum might be strengthened by treating literacy as a unitary construct:

This sense of integration of curriculum areas might be strengthened within the OSSLT by integrating reading and writing tasks. Students could use information from the reading passages to complete writing tasks and some of the writing tasks could be treated as reading tasks. (Peterson, 2003, p. 6)

Professor Suurtamm indicates that this concern applies to the Grade 9 assessment, as well:

An assessment that presents mathematics as isolated bits of content knowledge does not present students or teachers with a comprehensive picture of mathematics. In designing an assessment, it is important to focus on the important mathematics that is being assessed. Using an item mapping to curriculum expectations and Achievement Chart categories does not necessarily guarantee that the important mathematics in the curriculum is adequately assessed. (Suurtamm, Moisan, & Luthra, 2004a, p. 6)

Professor Lawson also worries about using the categories that appear in the curriculum’s Achievement Charts:

[T]he separation of mathematics processes into separate categories for measure is not defensible either psychometrically (these are not separate constructs … ) or cognitively. [It] does not reflect our present understanding of how students learn most effectively within reform instruction … (Lawson, 2003, p. 13)
In our analyses of the 2003 Grade 3 assessment, we found that, for reading, writing, and mathematics, the categories built into the test were detectable statistically in the results, but were very highly correlated. The high correlations mean that differences among categories should not be interpreted as evidence that students are stronger in one skill than in another. Similarly, our analyses of the data from the 2003 OSSLT provide some support for defining literacy as a single construct. The reading and writing measures are highly correlated. This is not to say that reading and writing are the same thing, but that the abilities corresponding to them nearly coincide in the student population, or that our measurements confound and intermix the constructs. The fact that the test currently requires the students to write answers to reading items and to read the instructions and prompts for the writing items may also account for the fact that, statistically, the reading and writing scores that are currently produced are correlated.

**CONCLUSION**

The strong consensus among those reviewing the assessments was that clear statements need to be developed about what each assessment is intended to measure. In the process of creating the statements from existing materials, EQAO may discover that some areas cannot be clarified without making additional decisions. If so, we encourage EQAO to include educators and other stakeholders in its decision-making. The papers commissioned for this review may also provide helpful guidance. In the end, EQAO will have created a series of documents, which describe and document the constructs to be measured at both general and detailed levels, with specification of the relative importance of sub-constructs. These documents should include the reason for and occasion of inclusion of each section. The documents should be revisited annually, and edited if necessary, so that they are kept current. Together with a clarification of the overall purposes of the assessment programs, EQAO will then have the needed foundation for further development of its assessments.
CHAPTER 4
DESIGNING THE ASSESSMENT PROGRAMS

The purposes of the assessment programs, discussed in Chapter 2, might be thought of as the “why” of assessment; the constructs, discussed in Chapter 3, as the “what.” In this chapter we examine the “how” of assessment— that is, the designs.

Each of the designs we describe in this chapter has advantages and disadvantages. Which design is best for each assessment program depends on the “why” and the “what.” It also depends on practical considerations, such as the costs of developing and marking items and of printing and shipping materials, as well as the impact on students, teachers, principals, and others of administering the assessments. It depends, too, on psychometric considerations, such as the reliability, validity, and year-to-year comparability of the resulting scores. We will make recommendations based on our understandings of the most important purposes of the assessments and what constructs they are intended to measure. However, EQAO will make the final decision about a design for each assessment and should do so only after the purposes and constructs have been clearly delineated.

Best practices in test design require a careful balance of content specifications and weightings, item formats and numbers, sampling of items and students, reliability and comparability, linking and equating over administrations, and utilisation of secure and released item pools. In our analysis, we consider variations in each assessment corresponding to different interpretations of the design priorities, focusing first on the strengths and weaknesses of the current EQAO designs.

_Finding 3._ Since 1996, the context for EQAO’s assessment programs has changed. Because each program was designed to fit a specific context, the designs may need to be revised to accommodate added purposes and refined construct definitions and curriculum connections.

_Recommendation 3._ That EQAO make decisions about the assessment programs’ designs after the purposes have been clarified, the constructs defined, and the curriculum connections specified. That EQAO plan to implement any changes over two to three years, to ensure that the quality of the assessment items and marking and scaling processes are maintained. That, for the individual assessment programs, EQAO make the following changes:

- For the Grade 3 and 6 assessments, move to a design that is modular rather than thematic. Reuse some modules across years to facilitate the equating of results.
- For the Grade 9 assessment, move to a two-part design with part to be marked by the mathematics classroom teacher and incorporated into student course grades and part to be marked centrally and provide information at the school, board, and provincial levels about trends in achievement in mathematics.
• For the OSSLT, if the construct definition for literacy supports a unitary literacy, move to a one-test, one-score, one-decision design; otherwise, if two constructs are clearly defined, separate it into two tests, one measuring reading and the other writing. Implement the change as soon as practicable.

Finding 4. The current assessment programs require a considerable commitment of time by students and by schools. Because most of the student work must be marked by teachers during special summer sessions, the marking costs are high. The information currently reported is detailed, but may not be optimal for its expanded uses. It may be that, with careful redesign, the tests can be shortened while providing information better targeted to the needs of schools and the Ministry.

Recommendation 4. That EQAO, once the purposes of the assessments have been clarified, tailor to these purposes its requirements concerning comprehensiveness and detail of the measurement, and concerning accuracy and comparability of scores and reports at the individual, school, board and provincial levels. That these requirements, along with considerations of cost and time, determine which design is selected.

DESIGN CONSIDERATIONS
Before we describe the designs, we should mention that several themes appear repeatedly. In particular, some of the designs show an increase in the number of multiple-choice items. In addition, many of the designs mention “embedded” field test and equating items. And the application of most designs requires the use of “scaled” scoring systems. We will explain each of these in turn.

Increased Use of Multiple-Choice Items
Professor Mark Reckase, a testing expert at Michigan State University with many years of experience at the testing company ACT, in a paper outlining recommendations for the Grade 3 and 6 assessments, described the relative advantages of different types of items, as follows:

There are two basic types of test item tools – those that require the generation of responses and those that allow the student to select a response from among a set of possible responses. The former type, constructed-response items, needs to be evaluated using human judgement unless the form of the generated response is very limited, such as a single number or a limited number of words. The selected-response items have the advantage that they can be scored very quickly by computer because the human judgements about the quality of response are made before the item is administered rather than after the response is generated.

Items that require generation of a response are appropriate when the generation skill itself is important. The specific type of information that is obtained must be worth the cost of making individual judgements of student work rather than making judgements of multiple
possible responses before the item is administered. Constructed-response items would seem to be most appropriate when there are many possible responses and the uniqueness of the response is a critical part of the information that is desired. However, the process of students generating responses is very time-consuming, and unless it is the generation of response itself that is important, constructed-response items provide inefficient ways of probing for cognitive skills and knowledge.

The items that require selection from a set of previously developed options, usually called multiple-choice items, can be used to probe for cognitive skills and knowledge for a large number of areas because of the flexibility of form and because each item acquires information efficiently. Multiple-choice items have been developed to assess many higher-order-thinking skills by using stimulus material such as descriptions of experiments, or competing hypotheses for a scientific result. (Reckase, 2004b, pp. 1-2)

Professor Reckase went on to explain that item types need to be selected to fit what is being measured:

> Item types are neither good or bad in themselves, they are more or less appropriate as a probe for particular types of knowledge and skills. Writing out a simple answer does not make it a high-level skill. Selecting a choice from among equally attractive options can be a very difficult evaluative skill. (Reckase, 2004b, p. 2)

We believe that Professor Reckase makes a compelling argument for including more multiple-choice items. In addition, multiple-choice items have the advantage of taking very little time to administer, so that more items can be administered without increasing the testing time. Increasing the number of items increases the amount of the curriculum that can be sampled. Additional benefits are that the marking costs for multiple-choice items are very small compared to the costs for marking constructed-response items and that multiple-choice questions are usually less memorable than constructed-response items and so can more easily be kept secure from one administration to another.

**Matrix Sampling**

Most assessment designs include some form of “matrix sampling.” Not all students take all the test items used in a given assessment. There may be a core of items that all students do take, and the core can be a large or small part of the total, depending on the test design. Beyond that core, the test items are divided into two or more parallel samples, each of which is administered to a separate sample of students. The administration is by random rotation, which means that assignment of item sample to student sample is done at random. Usually, the rotation is within each classroom, so test forms containing the different item samples are handed out sequentially around the room. Effectively, each item gets answered by a random sample of students, and each student takes a random sample of items. Further discussion of matrix sampling appears in Appendix C of this report.
EQAO currently uses variations of item sampling. In the Grade 3 and 6 assessments, the four multiple-choice test forms are distributed at random, one per student. In Grade 9 and for the OSSLT, the embedded field test and equating items are included in matrixed test forms that are distributed randomly.

The fundamental purpose of matrix sampling is to extend the number of items included in an assessment. There may be two reasons for wanting to do this. First, it permits the administration of “extra” items to representative samples of students for field testing and for providing equating as explained below; matrix sampling is already being used for this purpose by EQAO. A second purpose, not yet realised in EQAO’s assessments but part of some of the options proposed in this chapter, is to improve the content coverage of an assessment without increasing the testing time for individual students. Suppose that we need 100 test items to cover a content domain in depth. We might divide them into two forms of 50 items each and administer the forms using matrix sampling. Provicially, we would have statistical information from all 100 items. For a school, we could aggregate the information from both forms and also have statistical information for all 100 items. Only for the calculation of an individual student score would we be limited to 50 items.

We propose in some of our designs that matrix sampling be used to extend the content coverage and detail in a subject area. It is probably not appropriate when individual scores have high stakes, as in the OSSLT. But it is a promising methodology when the emphasis is on school, board, and provincial results, rather, than on individual student results.

Embedded Field Test and Equating Items
In Chapter 2, we described the importance of equating the assessment results between years before comparing them. We also discussed the difficulties presented by the design currently used by the Grade 3 and 6 assessments. If a clarification of the assessment programs’ purposes identifies year-to-year comparisons as a primary purpose, then the links between years need to be given careful consideration. We recommend that the links from one year’s assessment to the next should be based on equating modules embedded in the main assessments. This model is used now in parts of the Grade 9 assessment and in the reading portion of the OSSLT. The equating items should mirror the whole assessment in item types and content and should be kept secure between administrations. Students should not be able to easily distinguish which items will be used for equating.

Items that are being field tested – that is, administered, not to assess the students, but to find out whether the items work as expected – should also be administered with the full assessment. As we will discuss in Chapter 5, it is very important that we have good data about new items before they appear on a regular assessment and count in the results. Administering new items during the regular assessment is the best way to find out how they will work when they really count.
Tests, Test Forms, & Scaled Scores
Implicit in these designs are two other recommendations. In a paper for this review, Professor Reckase argued eloquently for the use of continuous reporting scales for all of the assessment programs (Reckase, 2004b). In particular, he suggested that (1) the assessments should present scores on a continuum and (2) the assessments should use an item response theory (IRT) scaling approach to place items and student scores on the continuum. We will describe the details of placing scores on a continuum in Chapter 6 and further illustrate the methodology in Appendix B. When discussing the designs in this chapter, we will assume the scores will be scaled.

The fundamental effect of IRT scaling on test design is that it separates the definition of the specific tests and test forms that students take in an assessment from the definition of the scaled scores and consequent levels and decisions that derive from the scores. Students can take partly different forms within a single assessment and through IRT scaling receive comparable scores. The results for students in one year can be reported in scores on the same IRT scale as the scores in a previous year and the same cut scores for levels and decisions can be used. *Note that, contrary to intuition and much historical practice, this is not logically possible when form-specific results, such as the percentage of items correct, are to be reported.* From the perspective of test design, IRT frees us from the need to use a single test or strictly parallel tests and allows us to focus on the quality of the information obtained in each test and test form. Of course, we also need to ensure that the test design provides adequate accuracy for each reporting purpose and sufficient linkage to maintain the common IRT scale.

**DESIGNS**
Each of the design options will be summarised by a graphic. The graphics use the following conventions:

- **Item Types**
  - Constructed-Response Items
  - Short-Answer Items
  - Multiple-Choice Items

- **Released and Secure Items**
  - ITEMS THAT WILL BE RELEASED
  - ITEMS THAT WILL REMAIN SECURE (AND BE REUSED) IN SUBSEQUENT YEARS
Possible Designs for the Grade 3 and Grade 6 Assessments of Reading, Writing, and Mathematics

Current Design

We described the current design for the Grade 3 and 6 assessments in some detail in Chapter 2 and in Appendix A. It has the advantage of mirroring good classroom assessment and instructional practices. However, the precision of year-to-year comparisons is limited because the equating is based solely on a multiple-choice booklet that is unlike the rest of the booklets. In addition, because the numbers of items contributing to student and school results are small, the results represent only a small sample of the curriculum.
Option 1: No Theme, Mixed Constructed-Response and Multiple-Choice Items

In Option 1, the assessments are no longer based on a theme. This makes the creation of test forms more flexible. Instead of multiple-choice items being presented in a separate booklet, constructed-response and multiple-choice items are presented throughout the test. For most of the testing time, all students receive the same items. However, there is also a part in which students receive one of several booklets that are assigned randomly. In the current design, four forms of the multiple-choice booklet are administered in just this way. The change, then, would be that the randomly-assigned forms in Option 1 would have both types of items. The “common” part of the test – the items that were administered to all students – would be released. The other part, which we call “matrixed” because different students receive different items (here we are referring to matrixing of items, not of students – all students would still sit the assessments), would be used for year-to-year equating. In fact, in subsequent years, some of the items in the matrixed portion could become the common portion and new items could be added to the matrixed portion.

This design would provide better year-to-year comparisons than the current design because the comparisons are based on both constructed-response and multiple-choice items. It also might provide somewhat better curriculum coverage, because it contains more multiple-choice items. However, it is not clear whether teachers would find this design as interesting as the current assessment as a basis for their own classroom assessments.
Option 2: Different Students Write Different Items (Item Matrixing)

Option 2 extends Option 1 by making more of the test matrixed and then by using the matrix portion to improve content validity. This design promises better curriculum coverage in reporting at the school level; school-level results may be more accurate and informative. In the above graphic, we show that the total array of items available for forming scores at the school (or board or provincial) level could be nine times greater than the array written by any one student. This provides essential improvement in the accuracy of measurement of the content domain. For example, reading can be evaluated over a greater number and variety of texts, mathematics problems can be presented with alternative contexts, and writing can be tested with different genres and prompts. This also may lead to additional detail in the measurement of subareas of the content domain at the school and board levels, but not for individual students.

From a statistical perspective, the number of items is being increased but the number of responses per student is held constant, so the number of student responses per item is reduced. For large aggregates, such as province as a whole, the reduced sample size for an item is not a problem. For small aggregates such as schools, there may not be a sufficient sample for a given matrixed form to report detailed results by item, but there is still greatly increased accuracy for the content domain and major parts when we pool and average over the matrix forms. That is, a school score on, say, reading would be based on all forms and all students.
In the graphic, we show the student results as being based on only the common items, while the school results are based on all the items. There is another possible way to use these data. If an IRT scaling approach were used to produce scores, it would be possible to produce comparable scores for students based on all the items each student answered, even though some of the items were different. This approach, however, might lead to apparent discrepancies at the individual level between performance as estimated on the total array and performance on the released test. The policies on release would determine feasibility here.

**Option 3: Only a Sample of Students Are Tested (Student Matrixing)**

In Option 3, only a sample of students sits the assessment. This is probably not feasible if a goal of the assessment is to produce school or even board results. To get reliable results for a school, one would likely have to test most of the students, and to get reliable results for a board, one would have to test maybe one or two hundred schools. That is, the sampling density would have to be so high that one might as well do the complete population.

But this is the design of choice if, given a major policy shift, there are no requirements for school and board reporting—that is, if the goal is to produce a *provincial* assessment. Again, we see that clarification of goals is essential to working out the assessment design.
The purposes of such a operation would be to describe achievement and achievement distributions across the province, perhaps with some general breakdowns (e.g., region, rural/urban, size of school). When this is repeated over years, information is built up on the trends in achievement. This is the model for some of the previous provincial reviews and also is exactly the sort of survey and reporting structure used in national and international studies such as the School Achievement Indicators Program (SAIP) and the Progress in International Reading Literacy Study (PIRLS). The difference between those research studies and an EQAO-based sample assessment is that EQAO could focus specifically on Ontario curriculum and on associated variables, such as curriculum implementation and teacher training, that is of interest in Ontario.

**Summary of Grade 3 and 6 Designs**

At the beginning of this chapter, we mentioned that, in addition to the purposes of an assessment, the costs must also be considered when selecting a design. Here we consider a few of the costs.

- **Development costs:** Options 2 and 3 would require the development of more items.

- **Materials costs:** Options 2 and 3 might incur additional printing costs because of the numbers of versions of each booklet; however, if the overall length of the test is shortened, the shipping costs will decrease. Also, in Option 3, many fewer sets of materials are required. Sample surveys might include 5,000 to 10,000 students. Note that all students from French-language schools would likely have to participate.

- **Administration costs:** In Option 2, the test could be shortened, requiring less time to administer. In Option 3, only a sample of the students would be tested.

- **Marking costs:** In Option 2, the cost of marking would be reduced if the test length were reduced. In Option 3, the cost of marking would be greatly reduced, because fewer students would participate.

In addition to the costs to develop and mark the assessments, we must also consider their psychometric characteristics:

- **Validity:** The validity of the scores – that is, how well the test scores reflect the construct the test is intended to measure – would be improved for school- and district-level results for Option 2 because more of the curriculum would be covered due to the matrix sampling of items, even if the test were shortened. Option 3 would also produce more valid scores because of the matrix sampling of items. However, because only a sample of students would take the test, results would not be available at the school or student levels.

- **Reliability:** The reliability of the scores – that is, their accuracy and consistency – would be improved in Options 1, 2, and 3 because the proportion of multiple-choice items would be increased. However, reliability is also influenced by test length, so shortening the assessment in Options 2 or 3 might offset this gain. The reliability of results for aggregates (schools, boards, the province) would be increased, although in Option 3, there would be only provincial results.
• Comparability: The comparability of the scores – that is, how well they can be compared across years – would be greatly improved in all three of the options because the equating would no longer be based on only multiple-choice items.

The three options just described have assumed that clarifying the definition of the constructs to be assessed would not result in a change of focus. However, it is possible that work on the construct would suggest that a “Language” score should be reported instead of Reading and Writing scores.

**Status**

EQAO has already begun to make changes in the design of the Grade 3 and Grade 6 assessments, moving toward Option 1. EQAO will include pieces that anticipate a change in design in 2003-2004. The common part of the 2003-2004 assessment will follow the current design – that is, it will be based on a theme and will contain only constructed-response items. However, instead of the multiple-choice booklets, some students will receive modules that are not tied to a theme, but include both multiple-choice and constructed-response items. If these modules work well, they will be used to construct the 2004-2005 assessment. In addition, EQAO may consider reducing the length of the test in 2004-2005 or 2005-2006.

**Possible Designs for the Grade 9 Assessment of Mathematics**

This section outlines several possible designs for the Grade 9 Assessment of Mathematics. How well these designs can meet EQAO’s purposes is also considered.

**Current Design**

Because the Grade 9 Assessment of Mathematics was developed to assess students’ mastery of the Grade 9 mathematics curriculum, different versions of the assessment have been developed for students taking each course (the students taking locally developed courses)

---

2. The possibilities for equating from 2003-2004 backwards to 2002-2003 and forward to 2004-2005 are, unfortunately, not immediately predictable without examination of the data that arise. For the backwards compatibility, one can repeat the multiple-choice based equating, and relive the limited inferences about change. Perhaps some of the experimental measurements included in 2002-2003 and 2003-2004 will allow better calibration and reporting of the levels in the historical time series. The forward comparisons would depend on new scaling and standards determination in 2003-2004, and on the adequacy of the experimental items to carry forward that information to future tracking.
are not tested by EQAO). In addition, because students study in English or in French, the assessment must be presented in both languages. Therefore, four versions are to be created for each administration:

- an English language version for students taking Academic mathematics courses,
- a French language version for students taking Academic mathematics courses,
- an English language version for students taking Applied mathematics courses, and
- a French language version for students taking Applied mathematics courses.

Some students take their Grade 9 mathematics course in the first semester, while others take it in the second semester, and some, in unsemestered schools, take a full-year course. The test was administered at the end of each semester. Therefore, in 2003, there were two test administrations, with four different versions of the test at each administration. Each form of the assessment consisted of:

- 6 tasks (multi-part problems requiring several extended constructed-response items)
- 10 short-answer items
- 24 multiple-choice items

Each short-answer item and each multiple-choice item covered one strand and one category. The constructed-response items associated with each task together provided information about one or two strands and all four categories.

EQAO built into the assessment the possibility of linking performance across test versions: 2 of 6 tasks, 4 of 10 short-answer items, and 12 of 24 multiple-choice items were common across all four test versions within an administration. The proportion of common items was designed to permit linking across versions, if that was ever desired. The English and French versions of the test have been assigned to separate development teams, so that, except for the specified common items, the items are unique to a language and program. The development teams follow parallel processes.

The items are organised into booklets, as follows:

- Booklet 1: Multiple-choice items (30 minutes)
- Booklet 2: Tasks (administered as Part I [Tasks 1-3] and Part II [Tasks 4-6], requiring 40 minutes for each part, for administration at different sittings)
- Booklet 3: Short-answer items (30 minutes)
- Booklet 4: Items of all three types being pilot tested, field tested, or used for equating across administration (30 minutes)

All students taking, for example, the English Academic version in January (Semester 1) of 2003, received identical Booklets 1, 2, and 3. Many different versions of Booklet 4 were produced, in order to collect data for items being pilot tested (tried out for the first time,
with a small number of students), field tested (tried out for the second time, with a larger number of students), or used for equating (to link the test administrations; these items are administered to an even larger number of students). The versions were administered randomly to students. Not every version was administered in every classroom, to limit the potential damage that could be caused by a local breach of security.

**Option 1: Each Booklet has Mixed Item Types**

The only change represented in Option 1 is that each booklet contains a mix of item types (the double-length booklet that previously contained two sessions of tasks and associated constructed-response items is replaced by two booklets, each with one session of mixed item types). This design has two advantages over the current design: A student who is absent for a day of testing will not miss all of one item type and the booklet that contains the equating items will not be as obviously different from the other booklets. The fact that, in the current version, students can identify the part of the test that does not contribute to their scores is cause for worry, because the accuracy of the equating between years depends on those items being answered with the same level of effort as the common items.
Option 2: Different Students Write Different Items (Item Matrixing)

In Option 2, the booklets contain mixed item types, as they did in Option 1. However, the ratio of common to matrixed items has changed. In Option 1, one booklet contained matrixed items and these were used for field testing and equating, but did not contribute to scores. In Option 2, each student writes two booklets of matrixed items and these items contribute to school results. The scores for individual students are based only on the two booklets of common items in this design. Notice that this design also represents a decrease in test length from five booklets to four booklets.

Option 3: Same as Option 2, But Only the Matrixed Items Are Returned to EQAO

Option 3 is the same as Option 2, except that only the matrixed items are returned to EQAO. The booklets containing the common items remain in the schools and teachers have the
option of marking them and counting them toward the students’ course marks – perhaps even using them as the course final exam.

**Summary of Grade 9 Designs**

The purposes of the Grade 9 assessment will largely determine which of the options are possible. Our recommendation, if the purposes permit, is Option 3. Here is a summary of the costs:

- Development costs: Options 2 and 3 would require the development of more items.
- Materials costs: Options 2 and 3 might incur additional printing costs because of the numbers of versions of each booklet; however, if the overall length of the test is shortened, the shipping costs will decrease. In Option 3, EQAO would not produce student results, which would mean lower printing and shipping costs.
- Administration costs: In Options 2 and 3, the test would be shortened, requiring less time to administer.
- Marking costs: In Option 2, the cost of marking would be reduced because the test would be shorter. In Option 3, the cost to EQAO of marking would be greatly reduced, because only two of the four booklets would be marked centrally. However, in Option 3, the teachers of Grade 9 mathematics courses would mark their students’ work. If these marks could be used in place of a final exam, then teachers might, on balance, save time because they would be provided the items and marking guides instead of spending time developing them.

Of course, in addition to the costs to develop and mark the assessments, we must also consider their psychometric characteristics:

- Validity: The validity of the scores would be improved for school- and district-level results for Option 2 because more of the curriculum would be covered due to the matrix sampling of items, even though the test would be shortened.
- Reliability: The reliability of school-level scores would slightly decrease in Option 3, because they would be based on only two booklets for each student. The decrease is slight because the total array of items contributing to the school-level score would not change.
- Comparability: The comparability of the scores would be improved in all three of the options because the booklets containing the equating items would be less easy for students to distinguish, so effort could be expected to be more comparable across booklets.

**Status**

EQAO has begun considering these options, but will require clarification about the purposes of the assessment. At the very least, we suggest they move to Option 1 for the 2004-2005 assessments.
Possible Designs for the Ontario Secondary School Literacy Test

Current Design

In 2003, the OSSLT consisted of four booklets. The test was designed to be administered over two half days, with students completing two booklets per day.

Students were asked to read 12 reading prompts. (There were also three to four field test reading prompts that did not count in the student marks.) The reading prompts were distributed roughly equally over the four booklets and were of the following types:

- Information (e.g., explanation, instruction, editorial/opinion, argument; 50% of the prompts were of this type)
- Graphic (e.g., graph, schedule, instructions, promotional design or layout; 25% were of this type)
- Narrative (e.g., story, narrative; 25% were of this type)

Each booklet contained one writing prompt. The test as a whole contained one each of the following four types of prompt:

- Summary
- News report
- Multi-paragraph opinion
- Informational paragraph

The reading test contained three types of item: multiple-choice, constructed-response, and extended constructed-response. Forty percent of the items were multiple-choice, 35% constructed-response, and 25% extended constructed-response. Reading items were designed to tap three types of reading skill: understanding directly stated ideas and information, understanding indirectly stated ideas and information, and makes connections between personal experience and the ideas and information in the reading selections. Also tapped were four reading strategies: vocabulary, syntax, organisation, and graphic features.
One of the test booklets contained three additional field test reading passages with questions. These were kept secure and used to supply material for subsequent applications, with the statistical calibrations obtained during the field trial. This is called an “embedded” trial.

The writing test consisted of four prompts of the types described above. There were no multiple-choice or shorter constructed-response items related to writing. The prompt types are the same between administrations, but the specific prompts and supporting text and graphics are used just once.

An important problem with the current design is that students who write the test and pass one part have to sit for the whole test again but attend only to the part they still need to pass. As discussed elsewhere, this leads to a fundamental problem of fairness. A critical problem is that there is no systematic equating of the writing score and decision and so, effectively, the standard for writing needs to be reset each administration.

**Option 1. Fewer Items**

Option 1 involves relatively small changes to the OSSLT. There would be nine reading passages and three writing prompts, so that the test would be shorter. There could be a better distribution of reading passage types and questions, corresponding with the construct definition. The genres of the writing prompts might be rotated from year to year, possibly from a longer list of possibilities.

*This option is attractive from a practical point of view, but has some intractable problems.*

For the reading portion of the test, the item mix would be changed so that there would be more multiple-choice items and better constructed-response items. Student scores would be derived using an IRT scaling method. The reliability of the current reading test is very high (.94), and even though this plan calls for a reduction from 12 to 9 passages, it is considered that with technical improvements in items and scaling, the reliability will be maintained at a sufficiently high level. This is enough of a shift in the reading test so that it is probably appropriate to realign the scoring scale to the pass/fail decision with a formal standard setting. That cut score could then be maintained across administrations because of the IRT score equating.

The problem is with writing, which it is assumed still needs a separate score and decision. This option does not include any way to improve the calibration of the writing items between administrations, so the standards would still have to be reset each year. Moreover,
the reliability of the writing score and pass/fail decision is already problematic and would become worse; the reduction from four to three prompts would increase the granularity of the scoring, so that changes of one point on one prompt would move more students between pass and fail.

Furthermore, this design option still has the problem of how to fairly treat students who must retake the entire test after passing one part.

**Option 2. Two Tests, Two Scores**

As we described in Chapter 3, the OSSLT’s current design of a single test yielding two scores creates issues of standardisation and fairness for students who have already received one passing score, but must retake the entire test. This option may make construct definition more straightforward, as it is clear that there are two constructs, reading and writing, and would definitely clarify the situation for students who need to retake only one of the parts.

Having only nine passages would mean a shorter test, perhaps closer to two hours. The plan would be to increase the number of multiple-choice comprehension items and to improve the constructed-response items. Passage choice and item development should be guided by the definition of the construct. Constructed-response items might be developed in such a way that, at least for some passages, there would be only one question.

As with Option 1, scaled scores would be produced and good calibration and equating procedures established. Also, as with Option 1, a standard setting would be required.
The standalone writing test as indicated here would include some improvements that were proposed during the research consultations of this project and we understand are under discussion at EQAO. These involve some construct shift and measurement innovations, so further substantive and technical analysis is needed. The first innovation is to include a short test of conventions using multiple-choice or fill-in-the-blank items. This test would provide additional curriculum coverage and permit markers of the writing selections to focus on the content without fearing that knowledge of conventions is not being valued. The second innovation is to rearrange the writing tasks. There would be a small number of short writing items that involve minimal reading except for the brief instructions and prompts. Then there would be two writing prompts similar to the current prompts, perhaps with a rotation of genres. There might be one longer writing sample, including writing process requirements, to tap a different aspect of the writing construct. Note that the substantive analysis required to justify these changes revolves around the definition of the writing construct, and the technical analysis concerns the accuracy of measurement obtained from these new item types and arrangements.

Again, a standard setting would be required. The reliability of the writing scores and decisions probably can be made to equal those in the current design, even though the number of larger writing samples is reduced, because the conventions test and short writing items provide additional information, as does a longer writing sample. However, note that this design does not have any mechanism for equating results from one administration to the next, so some way of repeating the standard setting would be needed each year.

**Option 3. One Test, One Decision**

In this option, the test would look similar to the current test: It would include reading passages with multiple-choice and short-answer items and writing prompts. It might also include the innovations described in Option 2: a conventions test (multiple-choice or fill-in-the-blank) and short writing items. The longer writing samples might be similar to those in the current test, with a selection or rotation of genres, or they might include an extended
writing task. The appropriateness of these enhancements to the current test depends on how the construct is defined.

Overall, the test would be shorter, with perhaps two main writing prompts and eight reading passages, plus the conventions test and short writing items. Each of the reading passages might have eight multiple-choice items plus one written response aimed at getting a measure of reading ability and/or showing writing skill, maybe scored for both. All items on the test would be scaled together using IRT. Perhaps sub-scores for reading and writing would be produced in the same metric. The test package would include a matrix-sampled set of field test and equating items that mirror most of the assessment, although not the longer writing questions.

As with previous options, there would need to be a new standard setting and a mechanism for maintaining the standards across administrations.

The key improvement in Option 3 is that only one score and one decision for “literacy” is obtained, so that (a) the operational problems of partial re-testing disappear, (b) calibration is feasible for the test as a whole and the IRT scale and decision point can be maintained over administrations, and (c) there is a possibility of “compensation” between reading and writing for students who excel in one area but not the other. The first two points are practical and compelling. However, the third point reveals that a change of this nature has to be made with very careful attention to the intent of the test and the definition of the construct. Is this to be a test of reading, writing, or reading and writing, or reading and writing in reading? That is a substantive and a policy question.

At this moment, two alternative plans for implementing a single-score/single-decision OSSLT are under discussion.

**Option 3A. One Test, One Decision, Fast-Track Adoption in 2004**

This option calls for an adoption of one test, one decision in the 2004 administration, using the existing test constructed for the two-score procedure, but applying a new, unified scaling of the test, together with field test items, and a new standard setting. For students who have already passed half the test, there would be separate administrations of the reading and writing parts of the test scored according to last year’s system, but for first-time test-takers, the decision would be one overall pass or fail. In 2005, a new blueprint would be used for the test, only the single-score/single-decision procedure would be available, and the scales and cut score from 2004 and 2005 would be equated. The risk here is that the 2004 and 2005 test would be difficult to align, and perhaps a second standard setting would be needed in 2005. Also, there would be very little time after the October 2004 administration to carry out the scaling and standard setting. (A standard setting of some kind is required for writing, even in the current design.)
Option 3B. One Test, One Decision, Adoption Starting in 2005

This option calls for the 2004 test to be run according to the 2003 scoring and decision procedures. Only the originally-planned 12 reading and 4 writing components would be used, and students would receive separate results in reading and writing. (Again, it is strongly recommended that there should be separate administrations of the reading and writing parts for students who have already passed half the test.) After the administration and its reporting, there would be time to carry out validation, scaling, and standard setting using the assessment and the information from new items that were field tested at the same time, so that a prototype form according to the new design could be constructed from the field test items. Obviously, the advantage of this plan is that there would be more time for the implementation of the new test. The major disadvantage of this plan is that there would be no use of the single-score/single-decision procedure in 2004 and, even in 2005, there could not be a full implementation, since there would be students who passed only half the test in 2004 and who would require a reading or writing score and decision. Likely, this could not be obtained from a combined test if the test really is to be shorter and more efficient.

Summary of OSSLT Designs

The discussions around purpose and construct will largely determine which of the options is chosen. We do not recommend Option 1. Our recommendation, if the purposes permit, is Option 3 and we recommend moving toward it as rapidly as practicable. We recommend that as long as there is a two-decision test, those who rewrite only one part of the test should receive a completely separate test. This leads to the rather awkward advice that until there is a unified single-score/single-decision OSSLT, there will have to be three versions of the test. Here is a summary of the costs associated with the various options.

- Development costs: All options involve additional costs since new items need to be developed.
- Materials costs: Options 1 and 3 would lessen materials costs somewhat, while Option 3 would increase printing costs.
- Administration costs: Option 3 presents a clear advantage in terms of administration costs and effort, although starting later with 3b compared with 3a. Option 2 would provide some improvement.
- Marking costs: None of the options provide large decreases in marking costs. However, as the proportion of multiple-choice items is increased in Options 1, 2, and 3, or the number of writing prompts is decreased, costs may decrease somewhat.

In our view, the OSSLT will provide more accurate, fair, meaningful and practical results if the measurement and decision for reading and writing are combined. The increased accuracy will be obtained because the reading and writing components of the test will combine to allow improved equating and reliability from one administration to the next. The improved fairness and meaningfulness will be obtained by a more comprehensive view of literacy and by a compensatory model for students to demonstrate their literacy competence. The practical improvement will derive from the elimination of the need for separate tests and lingering decisions when students pass only part of the requirement.
CONCLUSION
EQAO commissioned this review to consider changes and refinements to the designs of its assessment programs. In this chapter, we have described some of the alternatives. Our recommendations are as follows:

• For the Grade 3 and 6 assessments, we recommend that EQAO adopt a design that is modular rather than thematic. The modules should be small sets of items that represent the overall test design within a subject area (reading, writing, mathematics), including the array of contents and item types. Some modules should be reused between years to facilitate equating of the results between years. Redefinition of purposes might lead to further simplifications and economies, especially if only provincial results are needed.

• For the Grade 9 assessment, we recommend that EQAO consider a design that would recast it to reflect two separate purposes. One is to provide a consistent tool for measuring student achievement in mathematics in a form that can be marked by the mathematics classroom teacher and incorporated into student course grades. The second purpose is to provide information at the school, board, and provincial levels about trends in achievement in mathematics.

• For the OSSLT, if the construct definition for literacy supports a unitary literacy, then EQAO should move to a one-test, one-score, one-decision design. If two constructs are clearly defined, then the OSSLT should be separated into two tests, one measuring reading and the other writing. The change should be implemented as soon as practicable.

It is very important, however, that any decisions about designs should follow the clarification of the purposes, definitions of the constructs, and connections to the curriculum. We worry that decisions made hastily may not reflect the ultimate purposes and constructs. In addition, changing an assessment program’s design requires very careful developmental and scaling work. If the types of items change, then the new items must be developed and field tested. If the construct definitions are refined or the curriculum connections clarified, then EQAO will need to examine the comparability of scores from new assessments with previous ones and the alignment of scores from the new assessments with the performance levels. Trying to make design changes too hastily could compromise the quality of the test items and of the results. As several of the experts who advised us during this review pointed out, changing an assessment program can rarely be accomplished in fewer than three years.
Any teacher who has created a test will tell you that writing the items is often the most difficult part. Each item represents an opportunity for the students to demonstrate what they know and can do. If the items are confusing, some students will answer incorrectly, not because they lacked the required knowledge and skill, but because they misunderstood the question or were distracted by something irrelevant in the question. If the items relate to a small part of the curriculum, students will be able to demonstrate their mastery of only that small part. If the items are too easy, then students who can perform at a high level will not have a chance to demonstrate that. If the items are too hard, then the students who have only limited mastery will not be able to demonstrate what they do know.

For a large-scale assessment program, having good quality items is even more important. An item on the Grade 3 assessment is administered to 140,000 students across Ontario. Items on the OSSLT are administered to about 150,000 students. To ensure the best possible items, large-scale assessment programs establish systems that include not only writing the items, but also reviewing them for sensitivity and assumptions about cultural knowledge, reviewing their fit into the test blueprint, and administering them to groups of students to find out whether they work as planned. In this chapter, we will describe the processes that EQAO currently uses for its assessment programs and suggest how those processes might be improved.

In the best practices of large-scale assessment, it is recognised that item writing and item validation are crucial elements. Experiences over many testing programs offer some important lessons. First, the number of items actually needed invariably exceeds the estimates made during initial planning of an assessment, since the estimates often fail to consider long-term requirements, selection for quality and content, need for release, and breaches of security. Second, items are much easier to write in some areas than in others, but the test design and item blueprint must determine how the item bank is populated, not ease and convenience of item-writing. Third, some people are much faster and better at writing items than others and should be encouraged and invited back. Fourth, performance items need many iterations of small-scale tryouts, because it is very difficult to predict how students will react to an open-ended prompt, and small changes in prompts can produce large changes in response. Fifth, accurate classification of items to content and expectation is not easy, and classification should be replicated and validated empirically.

Many of the items on EQAO’s assessments are very good. As we will discuss in this chapter, the main challenge EQAO faces is one of systematising the item-development and archiving process to improve the efficiency of the process and to add further safeguards against threats to the validity and quality of the tests.
Finding 5. While many of the items on EQAO’s assessments are of good quality, the processes by which items are developed, reviewed, and tried out allows some items of lesser quality to appear on the tests.

Recommendation 5. That EQAO establish and follow a standard process for the development and verification of items across all assessment programs. That EQAO create and maintain a general-purpose item banking and documentation system to support this process.

THE PROCESS OF ITEM DEVELOPMENT
Developing the items for an assessment is a long process. For example, EQAO spent three years developing the Grade 3 and 6 assessments that were administered in 2003. In the first year, reading passages were selected and items were written. After reading passages were selected and during the item development process, possible cultural sensitivity was considered. In addition, items were reviewed for out-dated or unusual words and to make sure they were written at the appropriate reading level.

In the second year, the passages and items were pilot tested in Grade 4 and Grade 7 classes to gather information about how students would respond to the items. Based on not only the students’ responses, but also feedback from the teachers who administered the items and the markers who reviewed the responses, the items were revised.

In the fall of the third year, the set of passages and items were field tested in Grade 4 and Grade 7 classes. The field test administration was intended to mirror the real administration that would happen in Grade 3 and Grade 6 classes in May of the same academic year. Following the field test administration, EQAO staff created “trait scoring guides.” Teachers were hired to use these guides to mark the student work from the field test administration and make recommendations about the items and the scoring guides. Based on the students’ responses, as well as feedback from the teachers who administered the field test and the teachers who marked them, minor revisions were made to the items. After these revisions, a field test in three classrooms in different regions of the province was performed to see whether the revised tasks functioned as expected. The students’ responses were analysed by EQAO staff and additional minor revisions were made where necessary.

The item development process for the Grade 3 and 6 assessments illustrates the many steps involved. A classroom teacher may be able to write items and put them immediately on a test to be administered to his class, but the quality requirements for items on a large-scale assessment are such that the items must be tried out at least twice and subjected to numerous reviews before they can become part of an assessment.

The flowchart on the next page outlines the major steps in an item development process, which we will discuss in greater detail later in this chapter.
Drafting and Editing Items

Professor Reckase, in a paper for this review, described creating items as an art, a “particular journalistic or literary form that needs to be mastered just as poetry writing or producing instructional text” (Reckase, 2004b, p. 1). He has compared the multiple-choice item form as akin to that of a sonnet. Professor Reckase’s enthusiasm for writing items may be unusual, but his comparisons underscore the difficulty of creating good items. The teachers and EQAO staff members assigned to draft and edit items should receive training in item writing. They should be given feedback on their items. Those who develop the skill of writing good items should be encouraged to continue writing items for EQAO’s assessments. According to several of the advisors to the review, in many testing programs, a few very talented item-writers turn out to produce a very large proportion of the good, finally usable items.
Reviewing Items for Sensitivity and Accuracy

In describing how items are developed for the Grade 3 and 6 assessments, we mentioned that passages and items are reviewed for cultural sensitivity and for unusual or out-dated vocabulary. Such reviews are necessary because, if some of the items are upsetting or set in an unfamiliar context for some students, those students will not have the same opportunity as other students to demonstrate their knowledge and skills.

It is also important to review items for accuracy. Items being developed for the Grade 9 assessment are validated by mathematics teachers and university mathematicians. These experts should also review the marking guides to check that the marking will reflect the knowledge and skill the items are designed to measure.

Pilot and Field Tests

Although asking teachers and specialists to review the items is very important, it is not sufficient to ensure that items will perform as anticipated. Items must be administered to students. The comments from teachers and students about the items may contain valuable suggestions for improving the items. Examination of the details of student responses, including the selection of wrong alternatives for multiple-choice items and the varieties of correct and incorrect constructed answers, help ascertain whether an item is engaging the intended knowledge and skill of the students. However, marking and analysing the students’ responses is also very important. Does the marking capture the evidence about the intended knowledge and skill? Statistical analyses of the patterns of responses can identify items that are too easy or too hard. Although items that the majority of students get correct may serve as “warm-up” items, the psychometric advantages and disadvantages of including such items in the operational assessment must be closely examined (Dunn & Emenogu, 2004).

It is also possible to tell if items are differentially difficult for one group of students – for example, an item about farm animals might be very difficult for urban students. Also, if students who perform well on the rest of the items do not do well on a particular item, that item should be reviewed for possible sources of confusion.

Based on the results of a pilot or field test, the item developers may decide to revise an item. If the item is changed, the new version should be subjected to reviews for sensitivity and accuracy. The new version should be field tested again. Making a revision to an item and then putting it on the regular assessment without first trying it out with students may be tempting, but is very risky. Even skilled and experienced item developers are often surprised about how students interpret items.

Reviewing Item Classifications

There is one additional review that we did not include in the process diagram, but which should be done once for each item. Each assessment’s blueprint will describe how many items of which types should be developed based on specific parts of the curriculum. As Professor Suurtamm’s research showed, though, categorising items according to the classifications used in the blueprint can be difficult, and different educators may classify the same items differently. Items should be classified by their developers, but also reclassified independently by educators not involved in creating the items. Where there are discrepancies in the classifications, the fit of the items into the blueprint should be re-examined.
ITEM BANKING AND DOCUMENTATION SYSTEMS
Developing items requires rigorously following a set of complex procedures. Tracking an item’s status, the revisions that have been made, which versions have been reviewed and by whom, and how the item performed in pilot and field tests, is important. Two types of computerised systems can help keep track of this information: an item banking system and a data and documentation archiving system. The item banking system records information by item: the item’s classifications, its performance on pilot and field tests, its record of revisions and reviews. The archiving system records information that is not specific to any individual item: data files from the assessment administrations, along with documentation of the data management and data analysis processes that created those files.

FINDINGS
Earlier in this chapter, we described the item development process that has been used for the Grade 3 and 6 assessments. The general steps followed for the Grade 9 assessment and the OSSLT are similar, although the specifics are different. Also, for Grade 9 and the OSSLT, it has been possible to embed much of the field testing into the “live” tests. That is, students in a regular assessment receive some number of items-in-development that do not count for test results but that are being checked or calibrated for future use or are present for linking score scales between assessments. By using small samples of items per student, the extra response burden is minimised, and by doing the testing during actual, in contrast with ad hoc or out-of-grade circumstances, accurate and realistic item results are obtained. With the Grade 3 and 6 assessments, unfortunately, the theme-based and non-modular design has precluded substantial use of embedded field testing.

For the Grade 9 assessment, the field testing and pilot testing of the items has occurred at the same time as the regular assessment. One of the four booklets that students received contained items that were being tried for the first time (pilot tested), items that were being tried for the second time (field tested), and items that were used to equate the test results across years. While the other three booklets were the same for all students, this booklet came in many different versions, because there were dozens of items to be tried out.

For the OSSLT, reading items have also been field tested during the regular administration. Pilot testing of reading and writing questions was carried out in volunteer schools and classrooms. All OSSLT materials were reviewed by educators and experts to ensure that they were fair, appropriate and free of bias. They were also piloted with different groups of students to ensure fairness for all students.

Some of the current processes have been found, over time, to make tracking and controlling item development difficult. For example, for the Grade 3 and 6 assessments, the use of a theme has caused EQAO to develop the entire assessment at once, then pilot and field test it as a whole. Unfortunately, this limits the improvements that can be made in the items during the development process. If an item does not work well at the pilot test, it can be revised, but not removed or replaced, because no extra items were developed. More typical of large-scale assessment programs would be to develop up to three times as many items as
will be needed on the assessment, so that items that do not work can be dropped, and so that the test developers who assemble the final version will still have some choices among items to try to balance the content the test covers. For reading passages or complex tasks, more items should be developed than will be needed, so that the best can be kept. In addition, extra passages and tasks should be developed. It is our understanding that EQAO has already started to make this change. There are also improvements that could be made in the timing of activities. For example, converting the items for the Grade 3 and 6 assessments to desk-top publishing format before the pilot test limits the ability to edit them and even to file them electronically, as each item is now part of a larger document. It is our opinion that items should be stored individually and assembled into booklets only at the time of administration. If subsequent edits must be made, they should be made to the individual files, then transferred to the booklets. The current practice means that edits reside in the booklets, but nowhere else.

It is not clear to us whether items are always reviewed for sensitivity after each revision, and whether field test data are always examined for evidence of functioning across groups of students. These steps are important. We examined the English-language version of the 2003 Grade 3 Reading and Writing assessments for evidence of score differences by gender and status as an English language learner. We found that there is evidence of significant ESL and gender differences. The prevalence if not the magnitude of the ESL difference is probably underestimated since only students in formal ESL programs are identified, not the larger population of English language learners (in a paper for this review, Professor Jim Cummins noted that other research suggests that “on the average, a period of at least 5 years is typically required for ESL students to catch up in English literacy and other academic English abilities”). We found some differential effects by item type, which suggest that some items that are differentially difficult may be present. These results underscore the need for careful checking of each item’s text and its performance during its development. At the October 2003 Preliminary Research Results meeting, Professor Dany Laveault recommended that EQAO should be routinely doing Differential Item Functioning (DIF) analyses for ESL, gender, and program.

Beyond refining the current processes for each assessment program, EQAO should consider developing a standardised process that applies, with some variations, to all the programs. They should also develop a single item banking and archiving system. EQAO has created numerous documents describing parts of its test development process. The Grade 9 team has made considerable progress in developing an item banking system. Both EQAO and this review project have been developing archives of data and documentation. These initiatives are very encouraging.

**CONCLUSION**

In this chapter, we described the process of creating items for the assessment programs. While devoting a chapter to item development may seem to some to be too much emphasis on processes, we believe that such attention is appropriate. After all, even the best assessment design will falter if the items are not of high quality.
Small-scale pilots of new items will always be needed as part of the item-writing process. But we think that embedding pilot and field trial items into a main assessment assures high-quality and efficient item development information and should be adopted as the preferred methodology. Indeed, this may influence the overall design of an assessment. In the Grade 3 and 6 assessments, we suggest that the current use of a content theme may need to be dropped from the overall design to allow more flexible arrangements for embedding new items.

We believe EQAO would benefit from an organisation-wide understanding of and commitment to a test development process that incorporates the best practices of the testing industry. This process should be supported by an item banking system and an archiving system. The item banking system, in particular, can be developed or purchased. The Grade 9 assessment’s system has promise and might be developed into a general system. This will require a significant commitment of funds and hiring staff with specialised expertise. Not only must the system be extended, but also professional documentation must be created and formal training of users must be provided. Even after a system is created, maintaining it will require continuing effort and resources.
CHAPTER 6
GETTING FROM THE STUDENTS’ WORK TO SCORES

As any teacher will tell you, after the students sit an assessment, there is still much to do. The students’ work must be marked, and the marks must be communicated in a meaningful way to the students. Sometimes this means translating the marks to a percentage or to performance levels. For a single classroom of students, these processes can be handled by the teacher. For all Grade 3 students (or Grade 6 or 9 students or students taking the OSSLT), however, the marking and scaling of the students’ responses are enormous tasks. In addition, because more than one teacher will be involved in marking, the process must be carefully structured with very clear instructions. In Chapter 3, we described the importance of documenting the decisions that are made in creating an assessment so that those decisions can be discussed. The decisions made in creating the marking guides must be documented for the same reason. In addition, a great deal of detail must be included in the guides if all the participants in the process are to apply them in the same way. This is critically important if every student is to have the same chance to have his or her knowledge and skills reflected in the final marks.

The process of translating the marks to the results that will be reported back to the students, parents, teachers, schools, and boards must also be carefully documented. For a large-scale assessment, this process involves several steps. First, the marks must be combined. The resulting total scores must then be equated with results from previous years – that is, as we described in Chapter 2, the scores must be placed on a common scale, so that any changes due simply to differences in item difficulty between years will be removed, leaving only real changes in the scores. Finally, the results must be reported in a meaningful way – for the EQAO assessments, this means reporting the results by performance level (as described in the Assessment Scales) or, for the OSSLT, as a pass or fail. These steps involve both statistical analyses and input from educators, and can together be called scaling.

The use of item response theory methods to integrate the scoring, scaling, equating, and reporting level setting has become regarded as best and standard practice in large scale assessment. EQAO has used this technology extensively in the Grade 9 assessment and in the OSSLT reading test. We have examined and evaluated the different practices used by EQAO and present various empirical studies in the research reports produced for this review. Our research and consultation in this area has mainly concerned how EQAO can extend IRT methods to all assessments in a consistent way.

Finding 6. The EQAO assessment programs now use a variety of approaches to marking and scaling students’ work. These approaches, while adequate, may not be optimal for producing the best quality results from the students’ work.
Recommendation 6. That EQAO develop a unified system for marking and scaling students’ work. That the marking be based on item-specific marking guides that align with the nature of the anticipated responses, not the final reporting scale. That the statistical scaling use item response theory models – preferably the Rasch family of models. That alignment of the scale to the reporting scale be based on formal standard setting procedures. That EQAO rigorously document all of its procedures for marking and scaling the students’ work for each assessment program.

MARKING THE STUDENTS’ RESPONSES

Marking Guides
The marking process begins with the development of marking guides. Sometimes, as we described in Chapter 5, these marking guides are developed when the item is first created. Usually, marking guides are developed at the field test stage, so that they can be used in marking the students’ work from the field test.

The different assessment programs have in the past used different kinds of marking guides. The Grade 3 and 6 assessments originally required markers to assign a single score across several items simultaneously. In 2002 and 2003, markers instead considered each item individually. A trait scoring guide had been developed for each item. The descriptors in the scoring guide were based on EQAO’s Assessment Scales and on the curriculum expectation the item was intended to assess. Teams of educators selected examples of student work that illustrated performance at each of the performance levels described in EQAO’s Assessment Scales. These examples were used as “anchors” against which markers judged the work of other students. In addition to the Levels 1 to 4, markers could indicate that a student’s work could not be assigned a level; the following codes were available: NE1 (Not Enough Evidence for Level 1), Blank, Unintelligible, Refusal and Off Topic/Irrelevant.

For the Grade 9 assessment, EQAO staff members drafted a specific marking guide for each “item” within the task (in this context, “item” meant the part or parts of a task to which a single score was assigned; parts of the task did not correspond on a one-to-one basis with the strand and category scores that were based on the task). The marking guide described possible student responses to the item and assigned a code to each. The codes were ordered approximately by the level of knowledge and skill the response displayed (occasionally, adjacent codes were considered approximately equivalent). The number of codes in each marking guide was determined by the complexity of the item. Some items had as few as three codes, while others had as many as six. Codes were added for blank and unintelligible work.

At the beginning of the summer session, the teachers who had been hired to mark the Grade 9 assessment were carefully trained to use the marking guides. In particular, the following points were stressed:

- Each code had a qualitative description of student work (sometimes, examples of responses were also included); the student’s work was to be assigned the code with the description that best matched the work
• The codes were not equivalent to the four levels in the math curriculum's Achievement Chart
• The codes were not equivalent from one item to another
• Different codes within a marking guide could represent the same level of achievement

For the OSSLT, marking materials included detailed descriptive rubrics and a number of marked examples for all the constructed-response items as well as the writing samples.

Marking Processes
For the Grade 3, 6, and 9 assessments, the marking is done during the summer, when teachers can be hired for two- or three-week periods. For the OSSLT, the marking is done in November and December, after the October administration of the assessment.

In the summer of 2003, the marking for the Grade 9 assessment began with marker training. The markers were divided into groups of about 20. Each group was assigned to a classroom and had an experienced marker as a leader and trainer. Each group marked one set of items for the entire two- or three-week session: the short-answer items, the Part I constructed-response items (associated with Tasks 1-3), or the Part II constructed-response items (associated with Tasks 4-6). The equating items and the June pilot and field test items were also marked during the summer session. The January pilot and field test items had been marked in February 2003. The multiple-choice items were scored by machine.

An important consideration in any marking process is making sure that the markers are interpreting and applying the marking guides in the same ways. For the Grade 9 assessment, three approaches were used to ensure marker reliability:

• Group marking: In the group marking procedure, all markers assigned to a booklet simultaneously marked a single student’s responses. A new student’s responses were used for group marking each day.

• Reinsertion: Some booklets were also reinserted by EQAO staff, so that they were marked a second time. In 2003, the booklets to be reinserted were selected according to a pattern intended to provide data for a generalizability study – in particular, efforts were made to ensure that sets of booklets were marked by sets of raters.

• Remarkling: Unlike the reinsertion process, which sought to sample all types of raters, the remarking process targeted particular raters who on the previous days had assigned codes that were unusually high or unusually low, had unusually high or low productivity, or were identified by a group leader as providing anomalous responses on the group marking.

The marking processes for the Grade 3 and 6 assessments were similar to that for the Grade 9 assessment. For the OSSLT, the processes were slightly different due to the importance of making accurate pass-fail decisions. Four different markers each scored one of the four reading parts on the reading component, and four different markers each scored one of the
four writing tasks on the writing component. All student papers close to the pass score were marked twice. Student writing that did not meet the criterion passing level (Level 3 or 4) was also coded as to reasons for the low marks (e.g., no supporting details) to provide information for feedback to failing students.

**SCALING THE STUDENTS’ RESPONSES**
Recall that, in Chapter 4, we recommended using continuous reporting scales for all of the assessment programs. We suggested that an item response theory approach be used to create the continuous scales. In this section, we will describe some of the details of such an approach.

**Scaling Approaches**
The approach currently used to scale the Grade 3 and 6 assessment results is based on classical test theory, in which total scores are sums of item marks. As we just described, the constructed-response items were marked in 2003 using trait-based marking guides with levels ostensibly corresponding to the levels on the Assessment Scales used for reporting the results. After the marking, the numerical values of the marks assigned to the constructed-response items were summed (e.g., a Level 1 contributed a score of 1). Standardised scores were created from the sum of the constructed-response items and equated versions of the sum of the multiple-choice scores and were weighted .8 and .2, respectively. The weighting represents the intended influence of the constructed-response and multiple-choice items.

Using methods described in the next section, the range of scores was divided into sections corresponding to Levels 1 to 4. Very low total scores received a score of Not Enough Evidence for Level 1.

The scaling of the Grade 9 assessment is based on an IRT approach. An IRT approach, instead of summing item marks, creates a model relating items’ difficulties and students’ scores, where both the items and the students are scaled onto the same continuum (IRT approaches are described in detail in several of the research reports prepared for this review). Complex IRT models were used: a 3-parameter logistic model for the multiple-choice items and the graded response model for the constructed-response items.

Scaling of the OSSLT reading items in 2003 was straightforward. The multiple-choice and short-answer questions were given a weight of 2 if correct and 0 if incorrect. The constructed-response with explanation questions could be scored 0, 1 or 2. Since there were 100 items on the test, there was a maximum of 200 possible marks. The writing items were scored on a 0 to 4 scale, corresponding to the Assessment Scale. The marks on the 4 tasks were summed. The pattern as well as the sum of marks was examined to determine whether a student should be passed or failed.
EQUATING BETWEEN ADMINISTRATIONS

In Chapter 2 and Appendix A, we described the difficulties in the Grade 3 and 6 equating approaches. Even though EQAO is careful to build each year's assessment based upon the same blueprint as the previous year's assessment, to use consistent administration procedures, and, if possible, to apply consistent marking and scaling procedures, the results cannot be assumed to be comparable across years. In the preceding section, we mentioned that the overall total scores were divided into ranges corresponding to the four performance levels on the Assessment Scale. The original ranges were defined in the first year of each assessment. In each subsequent year, the ranges have been equated – that is, adjusted statistically – so that the cut scores were comparable from year to year. The equating was based on students' performance on the items that were administered in multiple years. For the Grade 3 and 6 assessments, the only items that were administered in more than one year were the multiple-choice items. Therefore, the equating has been based on the multiple-choice items.

The Grade 9 assessment uses a “common item” design to equate results from each new administration back to previous administrations within language and program (results are not equated between languages or between programs). Take, for example, the English Academic forms. In January 2001, the first form was administered. In June 2001, a second form was administered. The forms had no items in common; however, a random sample of students responded to common items in Booklet 4. Recall that Booklet 4 contains either equating items, items being pilot tested or items being field tested. The equating items in both January and June 2001 were one complete form that was divided into 30-minute blocks and included in the rotated Booklet 4 versions. These equating items provided the link between the January and June forms (a statistical approach called “the Stocking and Lord procedure” is used to establish the link – this is described in detail in the descriptions of the assessment programs that were prepared for this review). The equating items from 2001 were administered as operational items in January 2002. New equating items were introduced in 2002 and then were administered as operational items in January 2003. New equating items were again introduced in 2003. Altering the importance of the equating items from one year to another (non-operational to operational) is highly problematic and likely to compromise the accuracy of the equating. This equating design is of particular concern, because many of the students are aware that some of the items do not count, thereby compromising their motivation (Herbert, Dunn, & Luthra, 2004). Unfortunately, the current equating procedure assumes that students’ motivation is consistent across all of the items.

The OSSLT reading items are equated across years using a similar approach – that is, common items across different versions of the assessment are compared/equated. EQAO uses several statistical approaches to equate the scores across years and combines the results across the different methods to determine the final equating. Where the Grade 3, 6, and 9 assessments adjust the students’ scores to have the same meaning across years, the OSSLT instead adjusts the pass-fail cut score against which students’ scores are compared. Year-to-year comparability in OSSLT writing has been more difficult to determine since
there has been no common content over the years. Instead, small sessions have been held each year to judgementally align the current items with the standard for passing that was set in 2000.

**Alignment to Standards**

So far, we have discussed how the items are marked, how those marks are combined into a score, and how the scores are made comparable across years. The next step is to determine how the scores should be translated into a meaningful result for students and schools. For the Grade 3, 6, and 9 assessments, this means reporting the results in relation to the curriculum's Achievement Charts as they have been adapted for EQAO's Assessment Scales.

For the Grade 3 and 6 assessments, there was no scaling, equating, or alignment in the first year of application (1996-97). Instead, a marker examined a student's work on all the items measuring a strand or category at once and decided what level of performance would be reported. In fact, scaling, equating, and alignment were not possible for this approach (and were possible, but not optimal, when a multiple-choice booklet was added). The original approach assumed that each year's assessment would be exactly as difficult as the previous year's. As we discussed in the last section, that, unfortunately, is unlikely to be true.

In the original approach, the teachers marked directly onto the levels of the Assessment Scale, but for categories or strands. The overall results were determined by averaging across the category and strand results. This might be criticised because the *average* of levels is not exactly a level; in fact, we understand that originally there was no intention to produce an overall result. In subsequent years, the multiple-choice items have been used to align the overall scores to those original judgements.

The Grade 9 assessment uses a different approach, the Scoring Guide Alignment (SGA) procedure, to determine to what range of scores each performance level should correspond. In the first summer of marking and using the SGA approach, individual scorers were asked to judge, for each item they had been scoring, how the scores aligned with the performance levels. Judgements such as these were collected for the constructed-response items for each test form (i.e., each combination of program, language, and administration). The scorer judgements and the IRT analysis results were combined using an analysis approach that is described in detail in one of the research reports written for the review (Falenchuk, Childs, & Dunn, 2004). Because results from subsequent administrations are equated back to the January 2001 scale, the cut scores were set once. SGA data have been collected during subsequent scoring sessions to confirm the reliability of the approach and the stability of the alignment (for this review, we analysed the additional data and confirmed that the results of the process are comparable across years).

In 2000, an extensive standard-setting process was completed for the OSSLT. The reading cut score could be carried through from that occasion because there have been common items throughout all versions of the test. Where this cut score falls each year is determined by equating the score scales. Because writing prompts are easily remembered, prompts are
not repeated across forms of the assessment, so instead of equating the results between years, a small session is held each year to judge how the scores from that year's assessment align with the passing standard set the first year.

FINDINGS

For this review, we conducted focus groups with teachers who were marking the Grade 3 and 6 assessments during the summer of 2003 (Elgie, 2004). The teachers reported that the anchors are important parts of the marking training, although they felt that the anchors do not fully describe the range of performance at the middle levels (Levels 2 and 3). They also considered the publication and distribution of the exemplars after the assessment to be very valuable for classroom teachers.

Our analyses of reinsertion data collected at the marking sites suggest that markers are interpreting and applying the marking guides reliably (Dunn, Childs, Cleland, Pang, & Saunders, 2004). In fact, the major sources of unreliability were not the marking process or marker variability, but rather item variability and item by student interactions. That is, some students did well on some items and others on other items. What this means is that, if the number of items is relatively small, as it is with the performance testing that EQAO prefers, reliability of individual scores suffers. Replicating marking will not help; only increasing the number of quality items will. Results of an analysis of partial data from the OSSLT marking site (for October 2002) tell the same story: It is not so much marker variance that creates unreliability in student scores, rather it is item or task variance and the interactions between items or tasks and individuals (Zhang & Wolfe, 2004). No increase in marker training or consistency will solve that; instead, one needs more and/or better items.

For this review, we reanalysed the Grade 3 reading items using IRT and found it feasible to use IRT scaling to bring together the performance items and the multiple-choice items from that assessment (Dunn & Emenogu, 2004). This can have important benefits, as IRT is a better method for: (i) providing information on a continuously graduated scale, and (ii) linking results from one assessment to the next. However, we found that some of the items did not perform well, which may mean that they are measuring something different from the rest of the items, or are confusing or overly difficult or easy. It will be important to improve the quality of the items.

We also investigated the feasibility of using simpler IRT models (specifically, the Rasch family of models) to scale the Grade 9 assessment. Our analyses suggested that, not only was such a change feasible, but the simpler models had some important advantages in terms of the fit of the models to the data and the ease of use (Falenchuk & Mazumder, 2004).

As part of the review, we sought advice from Professor Mark Wilson, an expert in large-scale assessment from the University of California at Berkeley. In a two-day meeting, Professor Wilson outlined how a Rasch-based scaling approach could apply to each of the assessment programs (the details are provided in a separate report prepared for this review). Professor Wilson pointed out that a single approach would better allow training and resources to be
allocated across the assessment programs. Using a Rasch-based approach will require that EQAO rigorously screen out poorly-performing items before they appear on a regular assessment. To identify items that should be dropped, EQAO will need to field test items with at least 250 students (if there are common items between the French- and English-language versions, however, it may be possible to field test the French-language versions with as few as 100 students) and consider the analysis results carefully.

CONCLUSION
The interpretations of the EQAO results are intended to be criterion-referenced and based on the provincial standards. A formal alignment to the standards has never been formally done with the Grade 3 and 6 assessments, and in Grade 9 and the OSSLT there will be a need for new alignments as the designs are revised. We see this as a two-stage process.

First, student responses on items should be marked according to the kind and variety of evidence that is apparent in the responses with the goal of extracting just the information and detail that is actually encountered. Systems that depend on markers’ making extrapolated judgements that immediately evaluate and code a student’s work on one item to the full, final Assessment Scale are flawed for two reasons: (a) the full range of performance of the scale is not going to be seen in any single item, and (b) the connection to the scale will be lost in aggregation over items.

Second, the statistical methods of IRT should be used to combine the student responses, over multiple-choice and coded constructed-response items, into continuous numerical scales. Convenient numerical scales should be employed that reasonably indicate the accuracy of the information and can be maintained constantly from year to year. (Different scales should probably be used for different grades to avoid inappropriate connections and interpretations of growth.)

We recommend that, once scales are established: (1) the process of standard setting should be regarded as a formal and high-level function within EQAO, (2) the methodology of standard setting should be integrated with the IRT scaling of the tests, and (3) once standards are set, they should be held constant across years through their reference to cut-scores that are established with respect to the IRT scales. Periodically, when the test designs or the criterion bases for the levels undergo major shifts, the standard settings should be redone.

Different IRT applications are now used in some aspects of the EQAO assessment. These methods can provide appropriate solutions to critical aspects of calibrating and equating scores and supporting comparisons from across years. Their use should be extended to all the assessments. They also can provide important frameworks for displaying results and distributions. The Rasch model has some limitations and restrictions (some low-discriminating items will have to be rejected), but overall it is a very practical approach.
CHAPTER 7
REPORTING THE RESULTS

In the end, a large-scale assessment is worthwhile only if its results are interpretable and useful. The quality of the results depends, of course, on all the steps that came before. In earlier chapters, we described the difficult decisions that test designers must make about as they seek to obtain valid and reliable measures of student performance without requiring too much testing time. Often, the detail that educators would like for diagnosis of individual achievement and remediation or for curriculum evaluation and revision would require an unreasonably long test. A shorter test may be able to provide general information, but more results would be insufficiently accurate because of the heterogeneity of items and student responses to them and the relatively small sample of items that can be devoted to any specific area.

This presents a dilemma: To produce an accurate score for a whole domain – mathematics, for example – we need to sample items from across the domain. However, because the length of the test is limited, we will likely have only a few items from each of several subareas within the domain – for example, problem solving in geometry. The small number of items in each subdomain means that we cannot be confident about the representativeness of the items in the subdomain. Subdomain scores based on only a few items will not give us an accurate picture of a student’s mastery of that subdomain. Even when we combine the subdomain results across students to produce school, board, and provincial reports, subdomain scores based on small numbers of items will not provide reliable information for curriculum planning.3

Other assessment programs have faced this same dilemma and developed other ways to report scores that can increase their interpretability and utility. In this chapter, we describe how EQAO currently reports the results from its assessment programs and suggest alternative ways that information could be communicated.

Finding 7. The EQAO assessment programs have sought to produce detailed reports for students, schools, and boards. However, some of the results have been based on very few items. Such results are neither as reliable nor as representative of the strand or category as they should be to be truly useful to students, teachers, or school administrators.

Recommendation 7. That EQAO define specific requirements concerning comprehensiveness and detail of the measurement and concerning accuracy and comparability of scores and reports at the individual, school, board and provincial levels. That EQAO consider alternative approaches, especially item mapping, to reporting information and providing interpretation of scales.

3. A matrix-sampled design may permit increasing the number of items sufficiently to provide adequate representation of subdomains for aggregate reports.
LEVELS OF REPORTING
For each assessment program, EQAO reports results for individual students, for schools and boards, and for the province as a whole. For example, for the Grade 3 and 6 assessments, in October 2003, each student who had written the test in May 2003 received an Individual Student Report (ISR) which contained the student’s overall performance levels in reading, writing, and mathematics. Students’ average scores for subsets of items defined by the categories and strands were reported to schools and boards (but not published).

Also in October 2003, schools and boards received summary reports for their students (EQAO released these results to the public in November 2003). The school and board results were reported in two ways. Method 1 expressed the number of students achieving at each level as a percentage of all students in the grade, including students who were exempted and those whose work on the assessment did not produce enough information to score. Method 2 expressed the distribution of student results as a percentage of those students who actually took part in the assessment and produced work that could be scored. Students who were exempted were not included in Method 2.

School and board (and provincial) reports also showed the number of students achieving at each level broken down by contextual factors, including gender, English language learner status, and special needs status.

Board and school results were released only if the number of students in the school or board was large enough:

- A school report was suppressed if the school had fewer than 15 eligible students (in Method 1).
- A board report was suppressed if the board had fewer than 4 schools and 1 school was suppressed or the board had fewer than 15 eligible students (in Method 1).
- In both school and board reports, Method 1 and Method 2 statistics for a subject area were suppressed if the total number of eligible students (Method 1) was fewer than 15.
- In both school and board reports, Method 1 and Method 2 statistics for contextual factors were suppressed for performance levels where the number of students was fewer than four.

In October 2003, EQAO released a report summarising the performance of students across the province. The provincial report contained both the overall levels and the mean category and strand scores for each subject area. All the contextual breakdowns were also provided. Provincial, school and board reports were available on the EQAO website.

After reports describing the results of the Grade 3 and 6 assessments were released, all the booklets of students’ work were returned to the schools. The text of the items, the rubrics by which they were marked, and examples of student work were posted on the EQAO website.
Students who wrote the Grade 9 assessment in January or June 2003 received ISRs in November 2003. The report included nine scores: an overall score, four strand scores, and four category scores. The scores were reported on the Assessment Scales (based on the Achievement Charts in the curriculum documents), ranging from Below Level 1 to Level 4. Scores were interpolated to the nearest 1/10th of a level. The scores were shown graphically. For example, a student with an overall score of 3.1 would have a dark point at 3.1 on a number line. For the overall score, but not for the strand and category scores, a confidence band was printed around the point estimate. The band spanned minus to plus one standard error. The standard error was the average across students, not the specific standard error for that student. The school, board, and provincial results for the Grade 9 assessment were presented as the percentage of students performing within a level.

For the OSSLT, students who sat the assessment in October 2003 received ISRs in late March 2004. Students were notified whether they passed or failed separately for reading and writing. Those who failed writing received diagnostic feedback pinpointing areas of weakness in their writing. Diagnostic feedback was provided for reading derived from the student’s response patterns to the 100 items.

For the OSSLT, reporting at the school, board and provincial levels followed the same general principles as those described for the Grade 3 and 6 tests. Reports on small schools, boards or board or school subpopulations were suppressed. As well as information about the number and percentage of students at each performance level, information was provided on absent, deferred and exempt students. Results were broken down by gender, special needs status, program of study (applied or academic), retake status (first-time versus repeat test taker), and English language learner status. As well, detailed tabulation of the student questionnaire was presented at the school and board levels.

**SCORE RELIABILITY**

When we talk about the reliability of results, we mean how accurate and consistent they are. Reliability typically depends on the number of items that contribute to a result, the types of items, and how the marking is done. For EQAO’s assessment programs, the biggest difficulty is the number of items. Constructed-response items take time to administer. This means that an assessment consisting mostly of constructed-response items will not have very many items. Most likely it will have enough items to produce a reliable overall score. However, reporting subscores based on just a few of the items may not be supportable.

For this review, we analysed data from the 2003 administrations of the assessment programs. We found that the reliability of the overall results is high. However, for a score based on only half as many items, such as a subscore, the reliability would drop below the level generally considered necessary for reporting to individuals.

We also asked Dr. Lauress Wise, the president of the Human Resources Research Organization, to consider the requirements for subscores. He first explained why it is tempting to produce subscores for the OSSLT:
The primary focus of the OSSLT is to provide summary information on student achievement relative to established standards. This includes whether students “pass” relative to graduation standards and the distribution of students by achievement levels for each school. When students fail to pass, they and their teachers would naturally like to know in what areas they should work to improve. When schools fail to improve achievement levels, they also would like to know which specific topics to work on to maximize school level improvement. (Wise, 2004, p. 1)

He went on to describe the limitations of such scores when they are used to guide instruction:

Another use of diagnostic information is to compare performance in one area with performance in another area. In this case, the statistic of interest is the difference between subscores. The problem with this use is that while the individual scales may achieve reasonable reliability, they are also highly correlated. Differences between highly correlated scores show little variation. Most of the variance in the difference scores is due to measurement error. As a consequence these difference scores will not meet conventional reliability standards.

Another problem with diagnostic scores is the difficulty in equating these scores across different forms of the test. Overall, there is a sufficient number of questions to enable test developers to keep test difficulty relatively constant across forms. Equating procedures can then adjust for minor differences in difficulty that are found. With small sets of test questions and a primary emphasis on meeting content standards, there will be relatively greater variation in difficulty across test forms. There also will be fewer common (anchor) items to use in equating subscores across the different forms. As a consequence, it is much more difficult to establish stable scales for smaller sets of questions than it is for the test as a whole. (Wise, 2004, p. 2)

Dr. Wise concluded by recommending that diagnostic scores be based on at least 25 items.

**SCORE VALIDITY**

When we talk about an assessment’s validity, we mean how well it measures what it is intended to measure. This means that an assessment is not in and of itself valid – its validity exists in relation to how it will be interpreted and used.

EQAO has always been careful to emphasise the criterion-referenced objective of its reporting systems. This means that students’ assessment results are reported in terms of what the students know and can do, and whether they meet criterion levels of performance that are defined in the content areas (reading, writing, and mathematics). What EQAO has tried to avoid is norm-referenced reporting and interpretation, in which students are judged according to their relative ranking on a test.

Confusion has long been rampant in the definition and interpretation of educational test scores, with some people favouring ranking measures such as percentiles, others trying to interpret percentage marks without reference to the content of a particular test form, and still others trying to impose meaning on arbitrary grading scales.
The emphasis at EQAO is on content-based interpretation of its reports, so the links that are made to the Assessment Scales are essential. The publication of exemplar student work for each level on the reporting scale helps to explain these links.

**ITEM MAPPING**

The EQAO assessments take a good deal of student and teacher time, and the hope has been that there will be substantial detail in the results – in particular, that subscores will be available to provide accurate and comparable information about differential student achievement in subareas of the curriculum and detailed information about strengths and weaknesses in school, board, and provincial implementation of the curriculum. However, subscores must be based on many test items, with careful attention to scaling, reliability, and comparability. If subscores are not scaled, reliable, and comparable, then their use is suspect. Generally, in large-scale assessment programs, the amount of testing time that would be required to produce numerous subscores is prohibitive.

One reason for interest in subscores is that it is hoped that assessment results will be diagnostic and prescriptive. This is, in many cases, not a realistic expectation. Creating a subscore that could be used for differential individual diagnosis and action requires just as much analysis, development and validation and nearly as many items as a total score. Perhaps, with the kinds of matrix sampling designs described earlier, one could obtain subscores that would be accurate and interpretable at aggregate levels (school, board, province) for purposes of educational planning, but individual students’ results would still not be accurate enough to support planning tailored to the individual students.

Interest in subscores also reflects a need to get precise and meaningful information from the assessments. What mathematics can students performing at a particular level do? What kinds of reading activities can almost all students do? What can only the best readers do? What is a typical writing product from a student? These kinds of questions can be answered with careful presentation of real results tied to the scales and statistics of the assessment. Such reporting is generally called “item mapping” (see Appendix B for examples).

One essential component of item mapping is the publication of example items. Usually, the text of the item (obviously, a released item), typical student responses and their marks, and statistics about how many students obtained each mark are included. Such information can very usefully be connected to a continuous reporting scale, such as one based on IRT analyses. At a given point on the scale, one or more items are shown – with text, responses, and statistics. These items are linked to that scale point in the sense that students with that score will have a high probability (say, 75%) of a correct response on a multiple-choice item or of receiving a particular mark on a constructed-response item. The items arrayed along the scale help to define the meaning of the scale. Item maps can also be used in school, board, and provincial reporting. For example, for a particular item, the distribution of marks within the school can be presented, along with comparative figures for the board and province, and perhaps even information from previous years, if the item was administered previously. It has to be emphasised that an item is a single opportunity for the student to
INCOMPLETE STUDENT WORK

When we talk about reporting assessment results, we assume that the students have responded to the test items. However, many students omit or provide unmarkable responses to one or more items. EQAO’s assessment programs have adopted a variety of ways to handle student work that is incomplete.

For some students taking the Grade 3 and 6 assessments, it was either not possible or not appropriate to calculate and report an overall performance level. This was particularly true of students who did not attempt significant portions of the assessment. Special designations were applied to such students. In reading and mathematics, students who answered fewer than half the questions in any category or strand received an overall score of NEIS (Not Enough Information to Score) for that subject. In writing, students who completed only one writing task received a score of NEIS. Students who did not answer at least one multiple-choice item from each block of questions for a subject also received an overall score of NEIS in the subject. Any student with one or more days’ absence was typically scored NEIS. Students who were absent for the entire assessment or who returned all booklets blank were reported as “No Data.”

The Grade 9 assessment also does not report results for students who omit too many items. However, for those students who attempted most of the items, the items they omitted are ignored when computing their score. In other words, the unanswered questions are treated as though the student never saw them, so there is no penalty for failing to attempt to answer (although, as noted above, if they attempted too few, they receive an NEIS designation). Research performed for this review shows that, even ignoring the items that are not attempted, the students with lower response rates also tend to have lower scores (Emenogu, Pang, & Falenchuk, 2004).

The approach used with the Grade 9 assessment is not completely satisfactory. There may be biases in treating non-attempted items as though they were not presented. We do not know enough about why students do not respond; it is possible that nonresponse is related not only to achievement, but also to other factors, such as interest, confidence, or speed.

How EQAO treats omitted items is largely a policy decision and needs to be clarified. Are we interested in how many students have mastered an expectation, or in how many will choose to try to do items related to that expectation? In higher-stakes testing, how omitted items will be treated must be very clearly explained to students. Some treatments of incomplete student work penalise students for not attempting all items. Others allow students to achieve high scores by doing careful work on only a few of the items. Students need to understand how omitted items will be treated, so they can choose an appropriate test taking strategy.
TIMING OF THE RESULTS
During the summer of 2003, we observed focus groups with teachers who were marking the Grade 3 and 6 assessments (Elgie, 2004). The teachers felt that the long delay between the administration and the reporting of the results was problematic. Here are examples of what they said:

…it would be great if [the ISRs] arrived earlier in the year, September no later than October and get those books to those Grade 4 teachers, so those kids will learn what the levels are about.

By the time [the students] get it back they will have long since forgotten. I would love an opportunity to go and sit down with one of those children or all of the children in some form and say this is what you got on this test, this is how you could have done better and a chance to almost take it up. I know there are logistic problems with that. It’s difficult but it would make it more meaningful for the kids. I think they would try harder…

I do my school improvement plan in September. I don’t get these results until November, so I’m working off the previous year’s results.

In addition, communication of results and sharing of returned student work was a problem within some schools. Some teachers felt that the results were not used effectively.

Well, number one, we have to look at when the books arrive back at your school, are they put in the corner of the principal’s office or is there any interest by the parent or teacher? And there is very minimal, I may get one call every couple years, so what happens to those books? To me they are very valuable only to be used to a great extent with the Grade 4 teachers...

We asked similar questions of teachers who were marking the Grade 9 assessment during the summer of 2003 (Herbert, Dunn, & Luthra, 2004). For most aspects of the assessment program, only about half the teachers reported satisfaction. For example, half the teachers reported that the results are useful and half found the administration convenient. In general, the teachers did not see the assessment program as fitting well with teaching and learning of Grade 9 mathematics.

FINDINGS
Several themes emerge from our research and the advice from our consultants: Concerns about the reliability of scores based on small numbers of items, the treatment of incomplete student work, and the timing of the reporting of results.

In large-scale assessments, there is a fundamental trade-off between the possible detail of reports and their reliability. The reliability depends on the number of items and the quality of evidence available for providing detailed information. In some cases, EQAO has tried to
report a larger number of detailed scores than can be reliably extracted from the student tests. In the Grade 3 and 6 assessments, it was hoped that 4 categories could be separately reported from 20 items. In Grade 9, the standard errors of some of the subscores are unacceptably large. In the OSSLT, the number of differentiable skills that could be reported for remediation planning was probably exaggerated. In the most recent assessments, EQAO has begun moving toward reporting fewer scores, each based on more items.

There are two principal reasons that educators may want detailed results. One is for diagnosis of individual students’ weaknesses and strengths or for planning school-wide instructional strategies. Producing detailed results for individuals based on large-scale assessments is very costly and such information may be better obtained by teachers using classroom tests and evaluation methods. However, detailed results for schools, boards, and the province for purposes of planning instruction might be obtainable using matrix sampling assessment design, in which many test items are administered within a school, but no single student takes all the items. The second reason educators want detailed results is to help them better understand the overall results, and for this, item-mapping techniques are potentially very useful.

The EQAO reports of the assessment results are intended to be of immediate and practical value to students, parents, teachers, administrators, and public. There are two problems, however. With hundreds of thousands of students sitting the assessments, there are bound to be students who do not respond to all the items, so that EQAO will have to decide how to report results based on incomplete data. Since there are individual reports, the resolution of incomplete results has to be precise for each individual. EQAO has developed detailed and complex rules for reporting; these rules need to be clearly communicated to students and teachers.

The bigger problem is the timing of the reports. In focus groups with Grade 3, 6, and 9 teachers who were marking the EQAO assessments, we heard that the assessment results are not useful because their return is out of sync with the educational processes they are supposed to inform and influence. For example, in the Grade 3 and 6 assessments, the student-level results are returned when the students are in Grades 4 and 7, usually with different teachers and often in different schools. In Grade 9, the results are not useful for student course grades: They are returned as much as nine months after some students complete the course. In the OSSLT, information about which students passed the test and which failed needs to be very quickly available if student course schedules are to be altered, so that those who failed can receive remediation.

**CONCLUSION**

We understand the desire to get maximally detailed and useful information out of the EQAO assessments. However, without increasing the length of the tests, it is not possible to produce valid and reliable scores much below the level of overall scores. It has to be accepted that individual diagnosis is not a viable objective for this kind of large-scale assessment. With matrixed data collection, we do anticipate more detail and accuracy for school, board and provincial reports.
EQAO should consider enhancing its reports at all levels with item mapping. This will allow much clearer understanding and interpretation of the results, particularly what it is that students at different levels of overall achievement know and can do.

Perhaps the most important problem around reporting is its timing. There are serious concerns in the field about whether the assessment results are available when schools need them for planning. The over-arching purpose of the EQAO assessment programs is to improve educational quality. This objective is jeopardised if results are not available when they can be used. One way to address this problem is Option 3 for the Grade 9 assessment, as presented in Chapter 4: The parts of the test for individual results could be left in the classroom for teacher scoring and immediate use.
CHAPTER 8  
ENSURING FUTURE QUALITY

In most educational jurisdictions in Canada and around the world, large-scale assessment has become an accepted if often contentious feature of how the public, educational authorities, teachers, students, and parents judge the quality of education and the successes and failings of schools and schooling. The range of interests and requirements can be seen in the array of assessments carried out by EQAO and the many uses that are made of EQAO reports. Our concern in this report has been with the technical and substantive quality of the EQAO assessments. We have documented many successes and identified areas where improvements can be made. The purpose of this final chapter is to discuss educational assessment in Ontario more generally and ask how EQAO can strengthen its long-term planning and continue to ensure quality assessment.

Finding 8. The EQAO assessment programs have many strengths. By following the recommendations in this report, they can be improved. However, future reviews will be needed, and an expanded, on-going program of research and development should be planned. The long-term goals of EQAO should be examined.

Recommendation 8. That EQAO strengthen the ways it carries out research, obtains advice, and plans for the long term:

- That the assessment program be supported by an active program of validity research.

- That there be improved long-term technical planning for the EQAO assessment programs and systems.

- That EQAO have a regular program of obtaining reviews and advice. That there be a high-level technical advisory panel with a regular schedule of meetings.

- That EQAO determine the time frame for its assessment programs and for periodic reviews. That EQAO carry out a major review every 3 to 5 years.

- That EQAO adopt assessment designs that are sufficiently general that they provide useful bridges across curriculum changes.

- That EQAO collect critical information to provide longer-term trend data.
VALIDITY RESEARCH AGENDA

EQAO is charged with collecting and reporting Ontario educational assessment data, just as Statistics Canada is charged with providing statistical information about the Canadian population, resources, economy, society, and culture. Statistics Canada runs the census that determines electoral riding allocations. Some EQAO assessments have also acquired administrative functions, such as the use of the OSSLT to determine student eligibility to receive a high-school diploma.

Neither agency could accomplish its work satisfactorily without an adequate research base and an ongoing research program. There is no universal operating manual that describes how to design and run a large-scale assessment program. Each aspect of a program requires research: the definition of the goals of the program, the analysis of the content and measurement framework, the test design and specification, the preparation and validation of items, the construction and calibration of tests, the accuracy of the interpretations, and the uses of the reports. Some research requirements are embedded in particular aspects of the work, such as test construction. Others are more general and refer to the interface between EQAO’s work and other parts of the education system.

The main focus for research needs to be the validity of the interpretations and uses of the scores and reports. The sources of invalidity are under-representation of the construct and the presence of construct-irrelevant variance, and it is critically important to understand and control these sources. Validity evidence supporting the interpretations and uses of the scores and reports needs to be based on examinations of the content, the processes students employ in responding to the test items, the marking and scoring, generalizability, external factors, and consequences. Of particular importance in the EQAO programs are issues of score interpretation (e.g., what kinds of reading or writing skills are engaged and measured?) and bias (e.g., are the tests fair for all groups of students?). A conscientious program of research should be based on a comprehensive agenda concerning what needs to be investigated and confirmed or corrected.

Here are a few research topics, chosen to illustrate the kinds of validity research we consider appropriate and, in some cases, urgent. Some of these research issues, such as item bias, need to be investigated, not just once, but periodically.

- Possible biases associated with culture and language background in the scores for the Grade 3 reading test (or any other EQAO assessment); differential item functioning across groups of students.
- Student problem-solving ability as revealed by different kinds of mathematics items.
- Continuity of measurement in mathematics from Grades 3 to 6 to 9; the analysis of constancy and growth.
- Analysis of the match between the test standards and the curriculum statements, especially with regard to the proportions of items in each area and to the areas not tested using current methods.
• Effects of EQAO testing in Grades 3, 6, and 9 on teacher selection and persistence.
• Effects of EQAO testing on coverage of the curriculum and on use of class time for test preparation.
• Relationship of EQAO test results with other factors such as classroom marks and final grades; concurrent and divergent validity.
• Effect of the OSSLT on rates of student progress toward graduation and on dropout rates.
• Details of reliability and classification errors for the performance levels and for the OSSLT pass/fail distribution.
• Accuracy and validity of aggregate scores for schools, board, and the province.
• Perceptions and understandings of the school reports by teachers, administrators, and the public.
• Appropriateness and utility of the Grade 9 test for students taking the applied mathematics course.
• Viability of making more use of translation between the French and English item banks.

This list could be extended. Clearly, some way must be found to identify the priorities for research. Some of these research areas relate to the operation and management of the assessment program and others relate more to the outcomes of the assessments.

TECHNICAL ADVICE
The EQAO assessments and the methodologies that support them are extraordinarily complicated. EQAO is operating with state-of-the-art goals and requirements, including content analysis, item development, test design and analysis, reporting, and validations throughout the process. Designing and maintaining the assessment programs and performing the research needed to support the programs requires expert advice. We recommend that EQAO establish a high-level technical advisory panel with world-class testing experts, psychometricians, and evaluators. This group should meet periodically to stay familiar with the Ontario program and situation, to review and critique plans, to help set the research agenda, and to provide ideas from experiences in other places. This technical panel should complement the committees EQAO now relies on for stakeholder perspectives.

LONG-TERM PLANNING AND REVIEWS
The EQAO assessment programs have served some important but relatively short-term goals, such as stimulating interest in classroom assessment and determining schools’ reactions to changes in the provincial curricula. The EQAO programs have also ended up providing educational indexes that are now routinely used in Ontario for judging students and schools. The match among the assessment programs, their purposes, and the uses of their results needs to be reviewed periodically. Reviews at 3- to 5-year intervals are typical practice in other jurisdictions.
The revisions and improvements to the assessments should be accompanied by long-term (3- to 5-year) planning for what is needed to provide effective and efficient support. This includes: (a) technical and substantive documentation, including frameworks, bridge documents, and blueprints; (b) item development requirements and sources, detailed by the number of test items required by each part of the test design; (c) a schedule of pilots, field trials, and assessments, with specification of items, forms, and samples; (d) data analysis plans; (e) a schedule of regular operations, including production, distribution, collection, processing, marking, scoring and scaling, calibration, analysis, and reporting; (f) materials for educators and the public; and (g) plans for research studies.

Long-term planning is particularly important because the kinds of designs EQAO uses and the enhancements we are proposing cannot be invented or reinvented in a year. It should take two to three years to make any major changes in an assessment program.

**LONG-TERM GOALS**

While the EQAO testing is generally based on the Ontario curriculum, it is clear that the curriculum is an ever-shifting target. Of course, the evaluation should follow and not lead the curriculum goals. At the same time, it should be recognised that some curriculum shifts have more to do with changes in instructional implementation than with changes in fundamental educational goals. For example, one can change the curriculum specifications of how reading might be taught or distributed across the grades, but the evaluation of reading can be relatively consistent and robust against such changes. EQAO assessment designs should reference general developmental understandings where feasible.

Another possible purpose of assessment programs is to track longer-term trends. The National Assessment of Educational Progress, which is administered by the US Department of Education to a random sample of students, for example, has student achievement trend lines that stretch back more than 30 years. The TIMSS project has trend analyses that will soon span 10 years. If EQAO wishes to report such trends, it will need to engage in careful planning going beyond what is needed for year-to-year equating and comparability. In fact, this goal will be difficult to attain if assessments involve yearly data collections and reporting, as well as strict adherence to curriculum specifications. However, current debates about educational quality reveal the importance of having long-term data to determine, for example, whether students now know more mathematics or can write better compared to five, ten, or fifteen years ago.

**LOOKING AHEAD**

Why is all this important? EQAO has created world-class educational assessment programs. As our review found, these assessments have many strengths, but can also be improved. We applaud EQAO’s commitment to continuing to refine the assessments. Because the assessments have become a vital part of Ontario’s education system, we believe that the effort required to maintain the best possible assessment programs is amply justified.
RESEARCH REPORTS FOR THE EQAO ASSESSMENT REVIEW PROJECT


CONSULTATION PAPERS COMMISSIONED BY THE
EQAO ASSESSMENT REVIEW PROJECT


Suurtamm, C., Moisan, P., & Luthra, V. (2004a). Does the mathematics being measured by the Grade 9 Assessment of Mathematics represent the “important” mathematics to be learned in the Grade 9 curriculum? Paper commissioned by the EQAO Assessment Review Project.


OTHER REFERENCES


APPENDIX A
A TECHNICAL HISTORY OF THE GRADE 3 AND 6 ASSESSMENTS

The Grades 3 and 6 Assessments in Reading, Writing, and Mathematics have been annual events in Ontario education since the 1996-97 school year, when the Grade 3 assessment began. As mentioned in Chapter 2, the history of the structure of the assessment and difficulties in making compatible measurements over time serve as an illustration of the effects of competing objectives in large-scale assessment. The following technical history describes the assessment design and the problems of comparability.

The assessment for each grade has been related to a theme, “change” in Grade 3 and “perspectives” in Grade 6. Within these broad themes were specific themes each year. For example, for the English-language versions administered in 2003, the Grade 3 theme was “turtles” and the Grade 6 theme was “whales.” Each student in Grade 3 received the following materials associated with change and turtles:

- Student Magazine: A colourful magazine containing two reading articles: one fiction and one non-fiction. Illustrations and graphics in the magazine supported the text.
- Student Reading Booklet: A booklet containing items related to two reading passages from the student magazine.
- Student Writing Booklet: A booklet containing two writing prompts.
- Student Mathematics Booklet: A booklet containing three mathematics investigations and their associated items.

The limitation of this design is that it makes compatible measurement between years very difficult. From year to year, how students perform on an assessment may vary. Some of the variation is because this year’s students have a better or worse mastery of the test content than the students did the previous year – they are more or less skilled. However, it is inevitable that the different items and the test as a whole will be somewhat harder or somewhat easier. Even though the same test design is used, it is impossible to ensure equal difficulty across versions of an assessment. To measure real improvement or decline between years, we have to be able to distinguish how much of the change is due to the particular items and test and how much is due to real changes in students’ performances.

There is an old maxim, “If you want to measure change, don’t change the measure!” (Mullis, et al., 2003, p. i). Repeating an entire test is rarely practical, however. The alternative is to statistically adjust or “equate” the assessment across years.

To equate the tests, EQAO must have either (a) some of the same items in both years or (b) some of the same students. Option (b) is not possible because the students who sat the previous year’s assessment are now in Grade 4. Repeating items is difficult because of the
theme – inserting items into an assessment with a different theme would be very awkward – and the items are memorable, they have been published on the web, and the students’ test booklets have been sent back to the schools.

To provide items that could be repeated across years, another section was added to the Grade 3 and 6 assessments: A booklet of multiple-choice items that were not related to the theme. Students, therefore, also received:

- Student Multiple-Choice Booklet: A booklet containing two language sections and two mathematics sections, for a total of 80 multiple-choice items. The language items follow reading passages and some are reading comprehension items, while others measure knowledge of writing conventions.

There have been four versions of the Student Multiple-Choice Booklet each year. Each student received one of the four versions. Each year, three versions were kept secure and one was retired. The retired booklet was then replaced with a new booklet the following year.

Three-quarters of the multiple-choice items are the same from year to year, so the students’ scores on the multiple-choice can be statistically equated and “true” change measured. Of course, this is change in that aspect of reading that is measured by the multiple-choice test. It overlaps with what is measured by the performance part of the test and is a part of the overall construct that EQAO is trying to measure, but it is incomplete and partly irrelevant. For example, the multiple-choice items do not adequately measure students’ personal connections to reading, but may measure irrelevant student test-taking skills.

The crux of the problem is that the equating for the whole EQAO Grade 3 and 6 assessments has been based on the equating of the multiple-choice items. The multiple-choice items account for only 20% of the overall score on the assessment. However, changes in performance on the multiple-choice items are interpreted as real change and translated directly to the overall assessment. If, for example, students in 2003 did better on the multiple-choice items than they did in 2002, EQAO would adjust the overall scores on the other items so that all of the test booklets show that the students are performing better than last year by the same amount. If the multiple-choice scores show no change, then no changes will be apparent in the overall results.

Technically, the equating of overall scores based on the multiple-choice items seems to work well. The alternative was to report raw, unequated performance scores, and that would have introduced uncontrolled, random, and doubtless wrongly interpreted fluctuations from year to year. This is observed in the currently unequated category and strand scores produced in the Grade 3 and 6 assessment and has led to probably spurious changes of improvement or decline in subareas. (The discontinuity between the overall and the category and strand scores and trends is another troublesome issue.)
However, practically and substantively, the multiple-choice based equating may not work so well. The problem arises if the multiple-choice items happen to measure somewhat different knowledge and skills than the items on the rest of the test, which require students to construct their answers. What if students in fact did know more, have more skill generally in 2003 than they did in 2002? Imagine that their performance on the theme-based part of the test and in general reading essentially improved – that is, would have been seen to have improved if the previous year’s test could have been re-administered. Unless the students’ performance on the multiple-choice items also improved, the change on the theme-based part would be assumed to be due only to easier items, so the overall scores would be adjusted to remain the same as the previous year.
APPENDIX B
ILLUSTRATIONS OF ITEM RESPONSE THEORY
SCALING AND ITEM MAPPING

The illustrations in this Appendix are derived from the consultation paper, *The purposes and function of a score scale*, prepared for the Review Project by Mark Reckase (2004b).

**Proposed IRT Scale for EQAO Grade 3/6 Assessments**
The following diagram shows one way to set an item response theory (IRT) scale for the assessments.

![Diagram of IRT Scale for Grade 3 and Grade 6](image)

The numerical ranges and locations could have been set at any arbitrary values, but the ones shown in the diagram were chosen for convenience in presentation and for supporting reasonable interpretation. The number of scale score points is based on the likely amount of accuracy in the test scores. The ranges for Grades 3 and 6 do not overlap significantly, because there is really no sense of continuity between the measurements. The centers and widths of the scales would be set at the moment of original construction of the IRT scales. Also, the cut
scores separating the provincial performance levels would be set the first year by a formal standard setting. In subsequent years, the scale and the cut scores would remain constant, although the distribution of actual achievement would be expected to rise.

The proposed reporting scale is established on the underlying IRT theta scale for each grade and subject area. In Grade 3, the means are set to 30, and in Grade 6, the means are set to 60. For both grades, the standard deviations are set to 5. There are 31 score points in each case and for individuals these would be reported as whole numbers. For school-board-provincial reports, results would be reported to one decimal place.

Cut scores for reporting levels would be established separately for each grade and subject area through a formal standard setting and then brought forward from year to year by IRT-based equating.

**Item Mapping for Mathematics Constructed-Response Items**

The next illustration is a mock-up of how EQAO might present an item mapping for a constructed-response item. A 5- to 10-page document would likely be produced, showing the results for different response levels for one item and for different items. There might also be information for the province and for a board or a school, although the primary purpose of the mapping is to illustrate the kinds of performance expected for students at different points on the IRT scale. Here it is supposed that the item was marked on a 3-point scale, and an exemplar is given of student work receiving a 3. The results mapped at the bottom of the page show how likely it is for students at different score ranges, or provincial performance levels, to get 2 or 3 of the possible points on this question.
Item mapping for mathematics constructed response

Question testing communication in measurement. The response to this question was rated on a 3-point scale, where 3 means that the mathematics concepts and procedures are correct and that the communication is appropriate, 2 means correct mathematics but inadequate communication, 1 means incorrect mathematics, and 0 means no coherent response. This response provided a generally clear and precise explanation of the relation of the triangle and rectangle, using appropriate mathematical terms and symbols. It was rated 3.

![Mathematics constructed response example]

Probability of response 2 or 3 according to scale score and level

<table>
<thead>
<tr>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>15-23</td>
<td>24-30</td>
<td>31-37</td>
<td>38-45</td>
</tr>
<tr>
<td>.11</td>
<td>.37</td>
<td>.70</td>
<td>.91</td>
</tr>
</tbody>
</table>
Item Mapping for Multiple-Choice Mathematics Items

The next illustration is a mock-up of how EQAO might present an item mapping for a set of multiple-choice items. There would be additional items to demonstrate variability of performance within and between different content subdomains. The goal is to show what kinds of content and procedures are mastered by students at different points on the IRT scale. That is, this demonstrates the meaning of the IRT scale. The results mapped at the right of the page show how likely students are to get the correct response depending on their overall scale score range or provincial performance level. Often a .75 probability of correct response is taken to indicate mastery of a content area.
Number sense

During her holiday, Mei Hwa bought gas 4 times. She bought 32.6 L, 52.9 L, 44.3 L, and 47.8 L. How much gas did she buy altogether?
A) 236.2 L  B) 175.2 L  C) 177.6 L  D) 187.6 L

.14 .44 .76 .94

Brad baked 720 cookies. How many boxes would he need if he packed 12 cookies in each?
A) 8640  B) 60  C) 732  D) 708

.12 .39 .72 .92

The place value of the digit 4 in 64 820 is:
A) ones  B) tens  C) hundreds  D) thousands

.14 .44 .76 .93

Measurement

Which one of the following has the greatest volume?
A) 750 mL  B) 1 L  C) 1500 mL  D) 2 L

.11 .39 .72 .92

Geometry

Which of the following figures is a sphere?
A)  B)  C)  D)

.26 .63 .87 .97

These multiple-choice questions correspond to different areas in the mathematics curriculum. Each has one correct answer. On the right are the probabilities of correct response for students with different scale score ranges, corresponding to the provincial achievement levels. A probability of .75 is considered to indicate mastery.
APPENDIX C
ILLUSTRATION OF MATRIX-SAMPLED ASSESSMENT

The diagram at the end of this appendix shows a typical application of matrix sampling as used in large-scale assessments. There is a core form (a set of common items that all students take) and other items are divided into additional test forms, which are distributed at random by rotating their distribution within each classroom.4

There are several circumstances where matrix sampling is already in use by EQAO and the proposed designs make additional and more substantive uses.

**Item Tryouts**
A matrix application permits field and pilot testing during the course of a regular assessment. This is now done by EQAO in all assessments. The rotated forms contain the new items, which usually do not count in a student score or in school/board/provincial aggregates. We do not need to have very large samples of data for field and pilot test items, just enough students so that statistics can be calculated for the item response distributions. This means that many rotated forms can be introduced, limited only by the number of available students and the printing complexities. Having tryouts within a main assessment yields much better samples than could be obtained by using volunteer schools, and the results are more representative of how the items are likely to perform when they become part of the operational assessment, because the students are in the right grade, taking the items at the right time of year, and under real testing conditions. In this kind of application, there is no particular need for the rotated forms to contain the same types of items. For example, one field test form might consist of constructed-response items and another of multiple-choice items.

**Equating**
In the EQAO Grade 9 assessment and in the OSSLT reading test, matrix sampling is used to obtain statistical linkage. Items that are matrixed into equating booklets in one year are incorporated into the operational assessment of subsequent years. The OSSLT reading test, for example, has rotated forms that contain passages and items that do not count in one year but that then become the elements for the live test in a subsequent year. Item response theory methodology, and also some classical methods, can be used to calibrate the new test score distribution back to the scale of the test in the year when the items were originally matrixed. In effect, the matrixed items are common between administrations and provide the information needed for equating.

---

4. Technically, this is systematic stratified sampling with random start, which has a substantial advantage of balancing the distribution of forms in each classroom and a slight risk of introducing correlations because of alphabetic orders or seating patterns. The core and rotated forms may be bound together so that it will not be evident to the students that there is any matrixing. Or the rotated forms may appear simply to be part of an ordinary division of a test into parts. In some versions of matrix sampling, there is no core form.
For purposes of equating, it is important that the common items mirror the whole assessment. If they are restricted to one part of the content domain, the equating will be biased. As discussed in Chapter 2 and in Appendix A, this is a flaw in the way multiple-choice matrixing has been used in the Grade 3 and 6 assessments.

**Content Elaboration**

A third kind of application of matrix sampling holds promise for enhancing the information obtained in some of EQAO’s assessments. Here matrixing is used to expand the item array that is used in an assessment. That is, within a given amount of student testing time, matrix sampling allows more items to be used. For example, in the Third International Mathematics and Science Study (TIMSS) 1995 design, each student took 90 minutes of mathematics and science testing, but data were collected, across students, on 396 minutes of items. The matrixing resulted in better coverage of the content domain.

Matrix sampling in this kind of application involves using the information from the rotated forms as part of the live test, even though only a fraction of the students respond to each set of items and even though each student only responds to a fraction of the items. This is first of all justified statistically and psychometrically because of the random assignment of forms to students, so that responses for each form are obtained from a random sample of the population. If the forms are constructed randomly from the item pool, then there is further justification for constructing student scores that will be comparable even though students take different forms.

For both aggregate statistics (school, board, and provincial results) and individual scores, which might be reported or used for distributional or relational analysis, there is going to be a trade-off between the increased validity in covering the content domain and decreased precision in individual measurement. This has to be evaluated carefully for each specific item pool and test design.

Use of matrix sampling for content elaboration is one of our suggestions for design enhancements in all assessments except the OSSLT. The OSSLT is a high-stakes test and it is likely that legal and ethical considerations would require that all students be tested and marked on exactly the same items.
A MATRIX SAMPLED ASSESSMENT DESIGN

The item pool for the test is divided into a core part (C) and two randomly parallel parts (A and B). Each student takes C and either A or B. The assignment is made by rotation (alternately) in each classroom. Usually students are not aware of which form they are taking.

<table>
<thead>
<tr>
<th>C. Core test form taken by all students</th>
<th>A. Rotated form, taken by one half of the students</th>
<th>B. Rotated form, taken by the other half of the students</th>
</tr>
</thead>
</table>

Each student takes C+A or C+B. Since A and B are parallel, the test information is comparable for all students. With IRT methods, the scores can be put on the same scale as if each student had taken all items. The per student accuracy will be lower.

Each classroom has all students taking C and random halves taking A and B. The data may be aggregated yielding results comparable to what would have been obtained had all students taken both A and B. There will be some reduction in accuracy.

Items are either taken by all students or by a random sample of one half the students, so population statistics can be obtained for each item in the pool. Statistics for the items in the rotated forms will be less accurate.
APPENDIX D
GLOSSARY OF TERMS

The terms in this appendix are arranged in the following categories:

- Measurement concepts
- Item types and components
- Test design
- Item and test development
- Testing operation
- Measurement technology
- Assessment reporting

Measurement Concepts

Bridge document
Documentation that establishes the link between curriculum specifications and test specifications

Comparability
Equivalence, between different test forms and administrations, of scores and their interpretations and consequences

Construct
Skills and knowledge that a test is intended to measure

Fairness
Equal opportunity for test-takers from different groups or with different backgrounds to demonstrate what they know and can do; lack of bias

Generalizability
Accuracy and statistical stability considering a variety of sources of variation and error: different students, different tests, markers, markings, times, etc.

Reliability
Statistical stability of the test score, especially over repeated measurements or alternate forms

Validity
Accuracy with which a test score reflects the construct to be measured; comprehensive coverage of the intended content and minimisation of irrelevant content

Validity research
Proactive program of verification and documentation of validity, reliability, generalizability, fairness, and comparability
### Item Types and Components

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authentic measurement</td>
<td>Test item that invokes student activity that directly represents the processes and skills called for in the construct, especially in the form usually used in classroom instruction.</td>
</tr>
<tr>
<td>Constructed-response item</td>
<td>Test item that requires students to write out or perform a response (e.g., a performance item, short-answer item, essay).</td>
</tr>
<tr>
<td>Embedded assessment</td>
<td>Testing that is carried out as a typical classroom activity, not in a special testing environment.</td>
</tr>
<tr>
<td>Multiple-choice item</td>
<td>Test item that requires students to make and record a selection from a fixed set of choices.</td>
</tr>
<tr>
<td>Performance item</td>
<td>Test item that requires students to demonstrate knowledge and skill by carrying out an activity that leaves a record of the processes utilised; especially an item with an extensive written or observed response.</td>
</tr>
<tr>
<td>Prompt</td>
<td>Material and instruction given to a student to stimulate the response to an item, especially a writing question.</td>
</tr>
<tr>
<td>Short-answer item</td>
<td>Constructed-response item requiring a short answer, usually a few words or sentences.</td>
</tr>
</tbody>
</table>

### Test Design

<table>
<thead>
<tr>
<th>Test Design</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embedded field trial</td>
<td>Use of matrix sampling to try out items that are still being developed with samples of students; an adjunct to a regular assessment.</td>
</tr>
<tr>
<td>Embedded equating</td>
<td>Use of matrix sampling to arrange common items between administrations of a test to allow statistical linkage of the scales and results.</td>
</tr>
<tr>
<td>Matrix sampling</td>
<td>Testing arrangement in which different students take different selections of items (forms), with sampling to ensure that each student gets a representative sample of items and each item is given to a representative sample of students.</td>
</tr>
</tbody>
</table>
Rotated form | A form used in a matrix sampled assessment that is alternated with other forms as the forms are distributed within a classroom, in order to assure a random sample of students within the classroom; sometimes called a spiralled form
---|---
Test blueprint | Detailed definition of the structure of a test, including the numbers and types of items
Test form | Instantiation of a test blueprint, as a particular set of items to be administered as a test

**Item and Test Development**

Field test | Final tryout of items during their development; students’ responses are marked and the data subjected to statistical analysis
---|---
Item banking | Systematisation of documentation about all current and past items, including text, classifications, application histories, and statistics
Item validation | Combination of documentation and examination of evidence from content classification, statistical analysis, judgements, and reactions to items
Item writing | Development of stimulus material, questions, multiple-response alternatives, scoring methods, and item content and process classifications
Piloting | Preliminary tryout of new items, gathering student and teacher reactions, checking students’ comprehension and the quality of responses

**Testing Operation**

Anchors | Examples of student responses that illustrate the levels in the marking rubric
---|---
Exemplars | Student responses, real or synthetic, that illustrate the possible levels of student response, from low to high
Marking | Judgmental coding of student responses, usually by teachers following a rubric
Marker reliability | Consistency of markers in judging student responses
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marking rubric</td>
<td>Instructions to markers; more generally, definition of the correspondence of student responses to the levels of the construct being measured</td>
</tr>
<tr>
<td><strong>Measurement Technology</strong></td>
<td></td>
</tr>
<tr>
<td>Equating</td>
<td>Establishing the points on the score scale for one test that match the scores on another, similar test</td>
</tr>
<tr>
<td>Calibration</td>
<td>Establishing the correspondence between two tests that are closely related but not strictly parallel; also, applying scaling methods such as item response theory to order items by difficulty and students by performance</td>
</tr>
<tr>
<td>Classical item analysis</td>
<td>Analysis of tests and items based on correlations and regressions</td>
</tr>
<tr>
<td>Item parameters</td>
<td>Numerical characteristics of items in item response theory, usually expressing the difficulty and discrimination of an item</td>
</tr>
<tr>
<td>Item response theory (IRT)</td>
<td>Analysis of tests and items based on a model of test response that assumes a probabilistic connection between the characteristics of the items and the characteristics of the students</td>
</tr>
<tr>
<td>Person parameters</td>
<td>Numerical characteristics of students in item response theory, usually expressing performance in relation to item difficulties</td>
</tr>
<tr>
<td>Standard setting</td>
<td>Judgmental process of determining which scores on the test score scale correspond to performance levels as defined in educational standards</td>
</tr>
<tr>
<td>Score scale</td>
<td>Transformed version of the theta metric that puts the student (and item) measurements in a convenient numerical form that retains its meaning across test forms</td>
</tr>
<tr>
<td>Theta metric</td>
<td>Theoretical continuous underlying scale that in item response theory is considered to jointly locate items and students</td>
</tr>
<tr>
<td><strong>Assessment Reporting</strong></td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>Incomplete student work</td>
<td>Records from students that do not contain all elements of the assessment, because of absence, error, or inattention</td>
</tr>
<tr>
<td>ISR</td>
<td>EQAO acronym for individual student report – that is, the paper report with explanation and results that is returned to the student</td>
</tr>
<tr>
<td>Method 1, Method 2</td>
<td>EQAO terms for tabulations across levels of achievement for an aggregation (school, board, province) with (Method 1) or without (Method 2) students who did not fully participate in the assessment</td>
</tr>
<tr>
<td>Item mapping</td>
<td>Methods for displaying items and statistical results for items with reference to the content organisation and the IRT score scale; ways of explaining the meaning of the score scale and the meaning of assessment results</td>
</tr>
<tr>
<td>PBS</td>
<td>EQAO acronym for Provincial-Board-School report – that is, the array of percentages, tables, and graphs that provide aggregate information</td>
</tr>
</tbody>
</table>