Interpreting and Using Results from Provincial Tests and Assessments

A Support Document for Teachers, Administrators, and Consultants

Manitoba

Interpreting and Using Results from Provincial Tests and Assessments

A Support Document for Teachers, Administrators, and Consultants

2009
Manitoba Education

Manitoba Education Cataloguing in Publication Data

371.26097127 Interpreting and using results from provincial tests and assessments : a support document for teachers, administrators, and consultants

Includes bibliographical references.
ISBN-13: 978-0-7711-4375-5

1. Educational tests and measurements.
2. Educational tests and measurements—Manitoba.
3. Students—Evaluation. 4. Students—Rating of.
5. Academic achievement—Evaluation. I. Manitoba. Manitoba Education.

Manitoba Education
School Programs Division
Winnipeg, Manitoba, Canada

Any websites referenced in this document are subject to change.

Print copies of this resource can be purchased from the Manitoba Text Book Bureau (stock number 80545). Order online at <www.mtbb.mb.ca>.

This resource is also available on the Manitoba Education website at <www.edu.gov.mb.ca/k12/assess/publications.html>.

Ce document est disponible en français.

Contents

Introduction

Manitoba Education ("the Department") is committed to assessment policies and practices that support improved student learning. The assessment* of students' progress toward achieving learning outcomes provides crucial information that assists teachers, parents,** administrators, and the students themselves in planning for improvement. Assessment is an ongoing part of the learning process, and occurs prior to, during, and following instruction. Many techniques are used, including classroom observation and dialogue, portfolios, teacher-constructed tests, and large-scale assessments. As assessment in its many forms takes on a broader role in informing the instructional process, it is crucial that the way information from these various kinds of assessment is used matches the purpose and capacity of the given assessment strategy. (These issues are addressed in the Manitoba document *Rethinking Classroom Assessment with Purpose in Mind: Assessment* for *Learning, Assessment* as *Learning, Assessment* of *Learning*.)

Assessment data can be used for a multitude of purposes, such as to provide feedback to students, plan instruction, certify achievement, report to parents, evaluate programs, and allocate resources. The information gained through a variety of assessment methods can help teachers, students, and administrators converse about and focus on student success and plan for and monitor improvement. However, to use the information from any assessment, one must know the purpose for which the assessment was designed and account for reliability and validity factors related to the resulting data and its interpretation.

This document focuses on data arising from provincial assessment policies. **Its purpose is to support teachers, administrators, and consultants in interpreting and using data from provincial assessments as they plan instruction or evaluate school and divisional programming**. As provincial assessment policies evolve, sections will be added to this document to support the appropriate use of the data arising from them.

Note that reference is made to current provincial policies. As policies are subject to change, readers are advised to verify whether or not the policy statements remain valid and to seek clarification or guidance in cases of doubt or change.

---

* An explanation of some assessment terms is provided in Appendix 1.

** In this document, the terms "parent" and "parents" refer to both parents and guardians. The term "parents" is used with the recognition that in some cases only one parent may be involved in a child's education.

While this document focuses on data arising from provincial assessment policies, assessment information arising from other sources, especially daily classroom-based assessments, is of the highest importance in terms of guiding instruction and communicating achievement information. Many of the ideas expressed in this document are applicable to other such types of assessment.

Grade 3 Assessment Policy

Purpose

The primary purpose of the Grade 3 Assessment Policy is to provide information to parents regarding their children's foundation knowledge and skills in reading and numeracy at the beginning of Grade 3 and in reading in French for students in French Immersion at the beginning of Grade 4. With this information, teachers and parents converse, reflect, plan, and work together to support improved learning for the child.*

Description

The Grade 3 Assessment is a classroom-based approach to assessing foundation knowledge and skills of students. Provincial policy outlines the critical competencies to be assessed, and teachers select or develop a variety of strategies to assess students' performance. Information is gathered by the Grade 3 teacher (or by the Grade 4 teacher in the French Immersion Program), as well as the Grade 2 teacher or school specialists, as appropriate. Teachers use continua (provided by the Department in the policy document) for reading and numeracy to evaluate student performance.

As assessment is integrated with daily classroom activities, teachers exercise professional judgment, supported by the continua and curricular documents, when identifying the processes for gathering assessment information. The exercise of professional judgment means that the Department does not expect nor consider it appropriate for every student to go through a comprehensive assessment on each competency for the purpose of this policy. The province also does not supply or prescribe specific assessment tools or strategies. In many cases, teachers are able to prepare the required reports on students' performance levels based on their daily work and conversations with students. Teachers may also use strategies to engage their students in reflecting on their learning.

Data Collected

Schools report to each parent individual student assessment information for each identified competency in reading and numeracy, indicating whether the child

- needs ongoing help
- needs some help to meet expectations
- meets expectations

* The policy document *Grade 3 Assessment in Reading, "Lecture" and Numeracy and Grade 4 Assessment in French Immersion "Lecture"* can be found at <www.edu.gov.mb.ca/k12/assess/docs/gr3/index.html>.

Schools also record the total number of students performing at these levels and forward this to their division which aggregates the data and forwards it the Department (independent schools send aggregate data directly to the Department). The Department prepares a provincial report which presents the percentage of students across the province performing at each of the three levels in each competency based on individual teachers' evaluations of their students.

Using the Results

Teachers and Principals

Results from the assessment are useful as a focus for reflection and discussion among educators as to the learning needs of students.

- They assist in overall instructional planning through identifying achievement trends.
- They can help identify specific areas of instruction for individuals or small groups of children.
- They can help determine the need for specialist intervention to address specific learning needs.
- They provide useful information for decisions about the allocation of resources or the review of intervention programs.

Student performance reflects, in part, students' cumulative growth and achievement as a result of instruction over previous years. Therefore, it is appropriate to share the information with all staff in a school to foster collaborative efforts to improve student learning.

Parents

With an understanding of the child's current level of performance and learning strengths and needs in specific competency areas, parents may become more involved in helping to support the child's learning. The assessment results provide a basis for informed discussions between parents and teachers.

School and Divisional Administrators/Trustees

By examining the assessment data contained in provincial, divisional, and school reports, it is possible to

- identify general trends in students' levels of achievement in the identified competencies as determined through the strategies selected by the teacher
- use observed trends to guide discussions among teachers and administrators regarding the implications for improving student learning

The Grade 3 Assessment information may be incorporated into school and divisional plans and reports, provided it is clearly indicated that these results are used only to help guide the classroom teachers in their work with students and discussions with parents during the school year. Because the purpose of the Grade 3 Assessment is to inform subsequent instruction at the classroom level, using these results for other purposes must be done with caution. See the Cautions section below.

The Department/Other Agencies

Observed trends in students' performance seen through the Grade 3 Assessment Policy serve as a point of discussion between the Department and educators from which it may be possible to identify specific areas of need that can be addressed at the provincial level. These might include plans for professional learning, identification of intervention programming, review of curricula, and adapting instruction. For example, the provincial Early Numeracy and Early Literacy Grants arose as a result of the analysis of the Grade 3 Assessment results.

Cautions

Data arising from the Grade 3 Assessment is meant primarily for teacher use in communicating with parents and in making instructional decisions. The teachers carrying out the assessment are in the best position to make meaning of the data relative to the way and the context in which the assessment was done in the classroom. Meanwhile, aggregating and reporting data can generate useful discussions about early years education provided that those involved in the discussion recognize the limitations of this data and triangulate it with information from other sources. Grade 3 Assessment data is an important source of achievement information as it is generated by teachers. However, it should not be considered as an independent source for drawing important conclusions about aggregated student achievement such as comparisons between groups or trends over time. Rather, it should be used only as one of several sources.

Because each language has unique features that are not readily equivalent and curriculum, instruction, and assessment processes differ accordingly, comparisons among school programs are inadvisable.

Middle Years Assessment Policy

Purpose

The first purpose of the Middle Years Assessment Policy* is to enhance student learning and engagement through classroom-based assessment processes that build student awareness and confidence in their learning. The policy is designed so that teachers can apply formative assessment practices (assessment *for* and *as* learning) to these ends.

The second purpose of the policy is to gather summative information about Middle Years student achievement in key areas, called competencies, attained by the end of January. These key competencies are number sense, number skills, and engagement with school at Grade 7, and expository writing and reading comprehension at Grade 8. The audiences for this summative information are

- the teacher, student, and parent who can use the information directly and immediately toward decisions regarding learning at the individual student level
- the school, school division, and Department who can use aggregate data to monitor trends towards making more informed decisions regarding, for example, the provision of resources to support improved student learning

Description

The Middle Years Assessment is a classroom-based approach to assessing certain key competencies. Provincial policy outlines the competencies to be assessed, and teachers select or develop a variety of strategies to assess students' performance in ways that engage students in the process (assessment *for* and *as* learning). Teachers use achievement information gathered in this way (over the first several months of the school year) to determine each student's most recent level of achievement for reporting purposes (to parents and to the Department) as of the last two weeks of January. Teachers determine the levels by using mid-year performance descriptors, supported by student exemplars, contained in support documents provided by the Department.

---

* The policy document *Middle Years Assessment of Key Competencies in Mathematics, Reading Comprehension, Expository Writing, and Student Engagement* can be found at <www.edu.gov.mb.ca/k12/assess/docs/my_policy/index.html>.

This assessment is integrated with daily classroom learning activities, and teachers exercise professional judgment, supported by the support and curriculum documents, when identifying the processes for gathering assessment information. This is intended to take place over the first several months of the school year. This assessment process employs the use of descriptive feedback and student-involved goal setting and criteria building that help students progress in the competency areas and that finally yield a summative report for parents of each student's achievement as of the last two weeks of January based on the most recent, stable performance. Schools also report achievement data on a student-by-student basis to the Department. The Department provides summary divisional feedback to school divisions to assist them with using the data for analysis and reporting purposes.

This assessment applies to all students in Grade 7 and Grade 8 enrolled in school divisions and in funded independent schools regardless of individual programming that might be in place. Only in rare cases where students cannot be assessed (e.g., late arrival in the school from another province) may students be exempted.

Data Collected

Schools report individual student assessment information to each parent for the competencies within mathematics at Grade 7 and language arts at Grade 8 according to the following performance levels:

- Not meeting mid-Grade 7/8 level of performance
- Approaching mid-Grade 7/8 level of performance
- Meeting mid-Grade 7/8 level of performance

Comment boxes are included on the reports to parents for use by teachers and by students to communicate contextual information regarding achievement and plans for learning, such as learning goals.

Students at the "Meeting . . ." level have either met or exceeded the mid-grade achievement levels for the competency. Students who are "Approaching . . ." require some extra attention, support, or effort to meet end-of-grade learning outcomes. Students performing at the level "Not meeting. . ." at mid-grade will likely not meet end-of-grade learning outcomes without an intervention aimed at building important knowledge and skills in these competencies.

Some students' levels of mid-year performance will be below the lowest performance level on the reports (below "Not meeting . . ."), most often due to a cognitive disability or their status as an additional language learner for which an individualized learning plan is in place. In such a case, only the comment boxes on the reports are used to communicate with parents regarding the student's circumstances and progress relative to the competencies.

For competencies within student engagement (Grade 7), the reporting levels are

- Emerging (only sometimes demonstrates the behaviour)
- Developing (quite often demonstrates the behaviour)
- Established (consistently demonstrates the behaviour)

In cases where there is a significant degree of variability in a student's engagement with their learning (across or within subjects), the performance category *Inconsistent* is used.

Using the Results

Teachers and Principals

Consistent with the primary purpose of the policy, a key result is the gradual increase in the awareness of and responsibility for learning by students through the use of formative assessment practices. Results for the Grade 7 competency *student engagement* can be used as one important indicator of the effectiveness of the formative assessment practices associated with the policy. Other indicators include

- improvements in student achievement
- general improvement in classroom dynamics such as more regular attendance, less disruptive behaviour, more time on task, and greater cooperation among students

Immediate and ongoing dialogue between students and teachers regarding student learning, which formative assessment practices entail, helps in planning learning experiences that improve achievement.

Summative performance results (as reported to parents and to the Department) from the assessment are useful as a focus for reflection and discussion in the following ways:

- The reports to parents provide a further opportunity to engage parents in helping to ensure that their children develop key competencies.
- When aggregated, the results provide useful information for decisions about instructional planning, the allocation of resources, and the review of intervention programs.

Student performance reflects, in part, students' cumulative growth and achievement as a result of instruction over previous years. Therefore, it is appropriate to share and discuss aggregate information with all staff in a school to foster collaborative efforts to improve student learning.

Parents

With an understanding of the child's current level of performance in the key competencies, parents may become more involved in helping to support the child's learning. The assessment results also provide a basis for informed discussions between parents and teachers.

School and Divisional Administrators/Trustees

By examining the assessment data contained in provincial, divisional, and school reports, it is possible to

- identify general trends in students' levels of achievement in the identified competencies
- use observed trends to guide discussions among teachers and administrators regarding the implications for improving student learning

Middle Years Assessment information is to be incorporated into school and divisional plans and reports to help guide classroom teachers in their work with students and in discussions with parents during the school year. Because the purposes of the Middle Years Assessment are to promote formative assessment practices and to provide information for making instructional decisions, using these results for other purposes must be done with caution. See the Cautions section below.

The Department/Other Agencies

Observed trends in students' performance serve as a point of discussion between the Department and educators from which it may be possible to identify specific areas of need that can be addressed at the provincial level. These might include plans for professional learning, identification of intervention programming, and review of curricula.

Cautions

Middle Years Assessment data is generated and used by teachers to make instructional decisions. When aggregating data and studying trends, it is important to acknowledge that a range of assessment strategies are available to teachers, which may have implications for how the data should be interpreted.

Meanwhile, aggregating and reporting data can generate useful discussions about Middle Years education provided that those involved in the discussion triangulate it with information from other sources. Middle Years Assessment data is an important source of achievement information as it is generated by teachers. It is appropriately used in combination with other sources for drawing important conclusions about student achievement.

When comparing language groups, it is important to recognize that each language has unique features and curriculum, instruction, and assessment processes differ accordingly. Comparisons should be done with caution, especially relative to competencies in reading and writing.

Grade 12 Standards Tests

Purpose

Standards tests are a summative evaluation intended to provide pertinent information about each student's accumulated knowledge and skills at the end of instruction relative to student learning outcomes in provincial curriculum documents. This information assists in certifying achievement and is used by teachers for evaluating instructional strategies for improving student learning.

Description

A current description of standards tests and the testing program is available in the document *Policies and Procedures for Standards Tests* and in the information bulletins available in all Grade 12 schools and on the Internet.* Standards tests are as curriculum-congruent as possible, within the parameters of large-scale testing. A student's performance on a test is properly understood as a "snapshot" of knowledge and skill relative to the representative sample of curricular learning outcomes addressed in the test.

Standards tests are centrally developed by expert teachers in the subject area and pilot tested to ensure the questions and tasks fairly represent the intended learning outcomes (i.e., have content validity). The tests are administered toward the end of each school year or semester and are scored by teachers at the local level using scoring keys, scoring rubrics, and exemplars provided by the Department.

Schools report individual standards test marks to students immediately following local marking and are required to indicate the standards test mark on the students' report cards. Standards tests count for 30% of a student's final grade in the course.

* These documents can be found at <www.edu.gov.mb.ca/k12/assess/publications.html>.

Data Collected

After standards tests are locally marked, schools and/or school divisions report each student's results to the Department. The Department compiles these results and returns provincial, divisional, and school profiles of performance including

- mean score
- pass rate
- scores on parts of the tests associated with a particular question type or group of related curricular learning outcomes

The Department also makes data available to each jurisdiction (school division or independent school) containing each student's test results on each question. As well, the Department summarizes and reports, at the school, divisional, and provincial levels, student participation information, such as the number of students exempted and absent from the test.

Samples of students' tests (excluding Consumer Mathematics) from each jurisdiction are re-marked centrally by the Department following local marking. The Department reports to school jurisdictions on the accuracy and consistency of local marking to assist jurisdictions with evaluating their local marking processes. See Appendix 2 for interpretive information on these reports.

Using the Results

Teachers

Standards test results are one indicator of student achievement relative to curricular learning outcomes and standards. As such, they are appropriate to use to

- prepare student evaluations and reports
- plan for instruction by identifying curricular learning outcomes for which there were relative strengths and weaknesses in student performance
- provide a focus for discussions among teachers and administrators regarding the implications for improving student learning

Reports from the Department for mathematics standards tests include comments regarding common errors (at the provincial level) for supporting discussions regarding standards test results.

Because standards tests reflect curricular learning outcomes and marking is guided by common (provincial), curriculum-based evaluation criteria (such as rubrics) and supported through training by the Department, teachers may use standards tests results to determine the degree of alignment between their own evaluation standards and those of the curriculum for the learning outcomes addressed in the test.

Teachers may also observe trends over time in standards test results (trend analysis). Although a different test is administered on each testing occasion, each test is made to the same specifications, and every effort is made to ensure close equivalency in overall test difficulty of the different tests. Any change in test specifications would be indicated in information bulletins that are sent to schools each fall.

Comparison from group to group and with provincial results (comparative analysis) is appropriate provided that test administration and marking procedures are carefully followed.

Students/Parents

Standards tests provide pertinent information regarding accumulated knowledge and skill relative to a representative sample of learning outcomes reflected in the test. As a result, they are appropriately viewed and used as evidence of learning and for reporting or portfolio purposes. This is the case, provided that there are no impediments to a student being able to demonstrate his or her knowledge and skill on the test and that marking guidelines are carefully followed.

School and Divisional Administrators/Trustees

Student performance on standards tests reflects cumulative growth and achievement as a result of instruction over previous years as well as during the current school year. As valid and reliable measures of student achievement, aggregated information from standards tests may be shared with staff in a school to foster collaborative efforts to improve student learning using means appropriate for the local context. Ideal vehicles include school and division plans and reports through which evidence-based discussion regarding items such as resource allocation may be fostered.

The feedback the Department provides regarding the consistency of marking (the marking feedback report) should be used by local authorities to evaluate local marking and when interpreting test scores. For example, a jurisdiction might observe from this report that they tended to mark tests too severely, and further, determine that this tendency was most apparent on questions of a certain type or from a certain curricular area. This information is especially important for comparative analysis and trend analysis. It is also important to consider student participation rates when looking at data over time or comparatively—this information is also summarized by the Department and returned to jurisdictions.

Standards tests that are properly administered and scored are appropriate for use in

- program evaluation, when it is important to have comparable data before and after the implementation of, for example, an intervention program
- trend analysis—looking at results over time relative to provincial level results
- comparative analysis—looking at school or divisional results relative to divisional or provincial results to identify, for example, specific areas of relative curricular strength in order to capitalize on them (share instructional strategies, for example)

The Department/Other Agencies

Data from standards tests reflects student achievement relative to the sampled curricular learning outcomes and standards, and may be used for a variety of purposes where this type of information is required. This includes studying questions such as gender differences or regional variations in performance, patterns in and needs related to curriculum implementation, and identifying potential areas for professional learning and curriculum revision.

Cautions

Standards tests are administered and scored locally according to prescribed, specific procedures. If these procedures are not followed, the reliability and validity of results are compromised.

Standards tests represent only one measure of student achievement. While carefully designed to ensure curriculum congruence, fairness, and accessibility to all students appropriately placed in the course, not all students are able to perform at their best in a final examination situation. To accurately determine any individual student's academic achievement, a variety of assessment strategies must be used.

As standards tests are locally marked, often by students' own teachers, it is important to consider the marking feedback report provided by the Department (to superintendents of school divisions and principals of independent schools) when interpreting test score results to ensure that there were no significant marking errors or bias. When schools and divisions organize for marking, it is strongly encouraged that

- standards tests be marked centrally by a group of teachers who train and work together
- tests be distributed randomly among the teacher-markers for marking
- occasional reviews of marking accuracy be conducted during the marking process (e.g., by double marking and discussion regarding differences)
- marking be "blind" (identity of student not known to the marker)

When using standards test results to study achievement patterns over time, it is important to recognize a number of extenuating factors that can affect the mean performance of a class, school, or division. These include the following:

- *any variation in local student exemption policies and the rate of student absence from the test*—While policies and procedures for exemptions are set at the provincial level, implementation is handled at the school level. The intention of the exemption policy is to excuse those students from participation in a standards test who have special needs for which there is no adaptation available that would make the test fair. Student absences from a test compromise interpretability of aggregate results, especially if the absences are likely to be test-avoidance strategies.*
- *slight variability in test content and difficulty from one test to another*—A change in a group average of up to about five percentage points should not be viewed as significant as the various tests are not precisely equivalent in terms of content and difficulty. However, should a consistent increase (or decrease) over test administrations be noted (such as, three increases in a row of 3% to 5%), or if a local change (e.g., school) in performance goes counter to larger trends (e.g., provincial), then this likely represents a meaningful change in group performance.
- *natural variability among groups of students taking the test at a particular time*—Random, uncontrollable, "natural" variability in group performance should be considered when aggregating and interpreting results for groups of about 30 students or smaller. For example, the presence of even one student with an unusually high or low mark can significantly affect a small-group mean score in a way unrelated to changes (over time) or differences (between groups) in factors that might affect group performance. Further information regarding such "statistical outliers" is provided in Appendix 3.
- *systemic variability in groups of students taking the test at a particular time*—Changes in course offerings or policies and practice regarding student enrolment in courses and programs can have a significant effect on test results. For example, if a particular mathematics course is offered in a school one year and not the next, then this would affect enrolment patterns and the types of interpretations that can be made when making year-to-year performance comparisons for mathematics.

There are specialized statistical procedures to use when comparing mean scores for smaller groups such as classrooms of students. For studies of this nature, the help of a statistical consultant should be sought if the expertise is not available in your organization.

---

* The Department prepares reports for schools and school divisions on their local marking results for standards tests, including information on the number and rate of students excluded from reporting. This report is also based on information provided by schools, and includes items such as exemptions and student absences at the school, division, and provincial levels.

Appendix 1: Explanation of Terms

Assessment: see also **Evaluation**

- Assessment *for* learning occurs throughout the learning process. It refers to strategies designed to determine students' learning needs, to plan instruction, to provide descriptive feedback to students, and to monitor progress towards learning outcomes.
- Assessment *as* learning refers to strategies that engage students in goal-setting and self-monitoring.

Assessment *for* learning and assessment *as* learning fall into the category of **formative assessment**. They are designed to support and improve learning rather than to assign marks to students.

- Assessment *of* learning, also referred to as **summative assessment**, refers to assessments that are typically administered at the end of a unit or a prescribed period of time to determine the extent to which learning outcomes have been attained and to what degree. Summative assessment is often used for evaluative purposes—to assign marks to students, to report to parents, and to measure programming effectiveness.

Continua are descriptors generally presented in a chart format that show what learners know and can do at various stages of learning or development, be it in reading, writing, mathematics, etc. They can be used to

- track student growth in a continuous fashion
- enhance the reliability of monitoring student development
- assist in planning and differentiation
- set learning goals
- report to parents

Continua may be *developmental*, meaning they reflect observable developmental characteristics of a learning process. Other similar tools focus on certain content skills that reflect a prescribed or expected learning sequence set out in curriculum documents—these are referred to as *scope* and *sequence*.

Criterion-referenced evaluation: see **Evaluation**

Evaluation is a term often used synonymously with **assessment**. Regardless of the term used, it is important to distinguish between two kinds of purposes. One is the process of generating and gathering data in a variety of ways for descriptive purposes. The other is the process of using data to make *value-based decisions* about students, such as pass-fail decisions, the need for remediation, or conclusive statements regarding a student's level of achievement (e.g., report card grading). **Reliability** and **validity** are significant considerations relative to both of these purposes for evaluation.

Types of evaluation include

- **Norm-referenced evaluation:** when a student's performance is described with reference to his or her ranking among some population of students (class, age group, grade, etc.)
- **Criterion-referenced evaluation:** when a student's level of achievement is described relative to fixed **standards** of performance
- **Self-referenced evaluation:** when a student's level of achievement is described in relation to the student's own growth over time

Exemplars are samples of student work or documentation of processes used to clarify criteria for student assessment or to illustrate strategies, and to clarify communications regarding assessment.

Formative assessment: see **Assessment**

Generalizability is the degree to which it is appropriate to interpret assessment data beyond the context in which it was collected. Generalizing assessment data requires that certain procedures be in place during collection and interpretation. For example, one would not generalize the results from a test administered to a sample of students to all students in a school division unless the sample was a representative, random sample of students in the division. Likewise, one would not make generalizations about a student's reading achievement unless the assessments that were done were thorough in addressing all aspects of reading. Finally, generalizability relates directly to when and if data collected at different sites (such as classrooms) can be properly compared. The assessment procedures (what was assessed and how) and student inclusion criteria (who was assessed) must be common or generalizable across the sources of data if comparisons are to be made.

Norm-referenced evaluation: see **Evaluation**

Reliability addresses the questions of whether *enough* information about student knowledge or skill has been gathered to be credible, and whether the observations, results, or reported scores would be the same if the information were gathered in a different way or interpreted by a different person (teacher), that is, the *consistency* of the measure.

Crucial to the question of reliability are factors such as

- the design of assessment instruments
- the care put into marking
- the clarity of the scoring rubrics/criteria
- the quantity of assessment information gathered
- the variety of types of opportunities for students to demonstrate what they know and can do

Extending the notion to student self-assessment, students need to be adequately familiar with the criteria (what is being assessed or monitored and the performance level descriptors, if applicable) and have adequate practice and feedback on self-assessment if they are to self-report "reliably."

Rubrics are guidelines for assessment that state the criteria upon which a student's process, performance, or product will be assessed. They often provide detailed descriptions of the quality of work at various levels of proficiency. They may be developed in collaboration with students and are most helpful when accompanied by **exemplars**.

Self-referenced evaluation: see **Evaluation**

Standards, in the context of assessment and evaluation, are fixed levels of achievement described, for example, in curriculum documents (learning outcomes and standards).

Standards Tests is a term used in Manitoba to indicate that our provincial tests at Grade 12 are based on curricular criteria (the learning outcomes) and **standards**. These tests are **summative** and **criterion-referenced**.

Standardized tests or **assessments** are instruments designed, administered, and scored according to set procedures so that, to the maximum extent possible, comparisons between students (a comparative evaluation) and comparisons over time (a trend analysis) are possible. Standardized tests are often **norm-referenced**.

Summative assessment: see **Assessment**

Validity addresses the question of *what* the assessment measures and *how* the information is used. While assessment data may be very reliable (precise and thorough), they may **not** be valid if

- the instruments or strategies used to collect the data (including marking guidelines or criteria) do not accurately reflect the knowledge and skill intended to be assessed (*content validity*) as described in curriculum documents
- the strategy used does not tap uniquely into the thinking processes or skills intended to be assessed (*construct validity*), for example, assessing reading comprehension through an extended writing piece
- the information is used in a way that does not suit the purpose or nature of the assessment resulting in unintended consequences (*consequential validity*)

Appendix 2: Marking Accuracy and Consistency

Grade 12 provincial tests are marked locally by teachers who use marking guidelines provided by the Department. Samples of test booklets are re-marked centrally by the Department. Comparisons of local and central marking are then done to generate reports on the *accuracy* and *consistency* of each jurisdiction's local marking. **It is important that the results in the *Marking Feedback Reports* be considered when studying Grade 12 provincial test performance trends in your jurisdiction.**

Two kinds of information about local marking are presented in the reports: marking *accuracy* and marking *consistency*. *Accuracy* refers to whether your jurisdiction's sampled test booklets score average would probably be the same when your local marking is compared to central marking, or if there is a bias in one direction or the other. *Accuracy* is evaluated through a common statistical test, called a T-test, and is tabulated in a column entitled Marking Accuracy in a table provided in the report. The T-test compares local marking and central marking of the sampled test booklets in terms of the test score differences. If there is a consistent pattern of local test scores being either higher or lower than central test scores, then the result of the T-test will be *statistically significant*, meaning that what has been observed in the sample is most likely true of the population from which the sample was selected.

Following are examples of these tables and their interpretation.

Example 1: Marking Accuracy: Average Test Score

Number of Students Writing	Sample Size	Test Score: Sample Means		Marking Accuracy
		Local Marking	Central Marking	
55	7	63.3%	61.2%	=

Analysis is based on a sample of seven tests. The equal sign (=) indicates that there is no statistically significant evidence of local marking being either consistently generous or stringent. Local marking was either consistently accurate, or inconsistent with some test booklets marked accurately, some generously, and some stringently. Marking *consistency* is addressed in detail further below.

Example 2: Marking Accuracy: Average Test Score

Number of Students Writing	Sample Size	Test Score: Sample Means		Marking Accuracy
		Local Marking	Central Marking	
55	7	62.7%	58.6%	+

The plus sign (+) indicates that there was a statistically significant tendency for local markers to be generous. Though local marking was *inaccurate*, it may still have been *consistent* (consistently generous, in this case), in which case the problem likely involves a marking tendency that can be identified and addressed, such as a particular interpretation of a scoring key or rubric.

Example 3: Marking Accuracy: Average Test Score

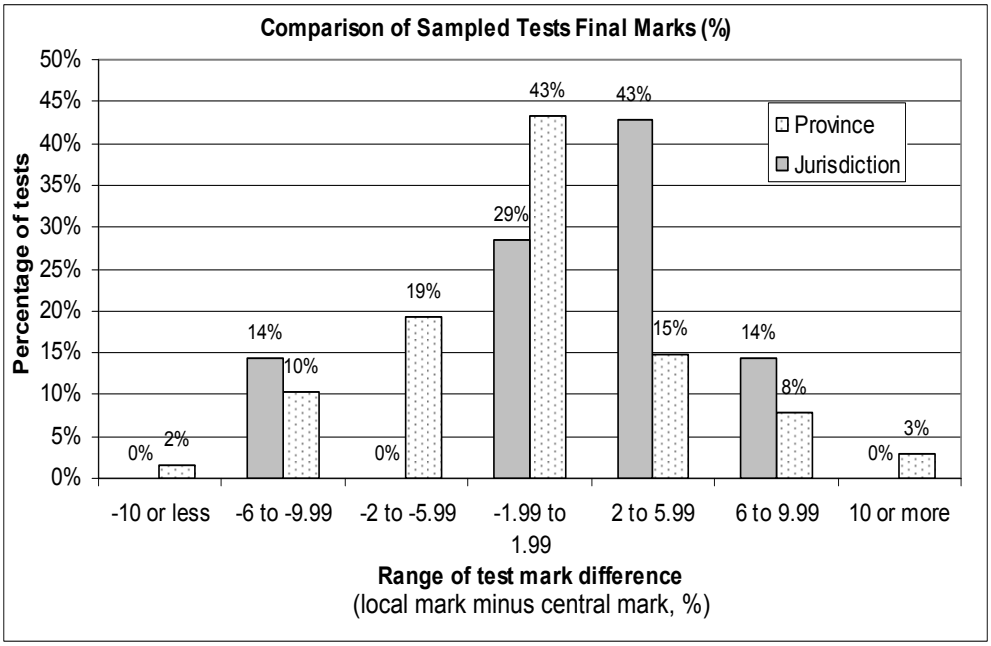
Number of Students Writing	Sample Size	Test Score: Sample Means		Marking Accuracy
		Local Marking	Central Marking	
357	22	64.4%	66.2%	-

The minus sign (-) indicates a statistically significant tendency for local markers to be stringent, albeit by a small amount (64.4 – 66.2 = 1.8%) on average. The larger sample size* (22) means the statistical test (T-test) is sensitive to even small degrees of bias. The fact that the difference (1.8%) is quite small **and** statistically significant implies that local marking was quite consistent.

* Higher sample sizes occur when necessary to ensure the sample is representative of a large number of schools and/or classrooms.

Marking *consistency* refers to whether markers are consistent in how they mark, be it accurate, generous, or stringent. Marking *consistency* is presented graphically in the marking feedback report, using a bar chart, as illustrated in the examples that follow.

Example 1: Marking Consistency: Test Scores



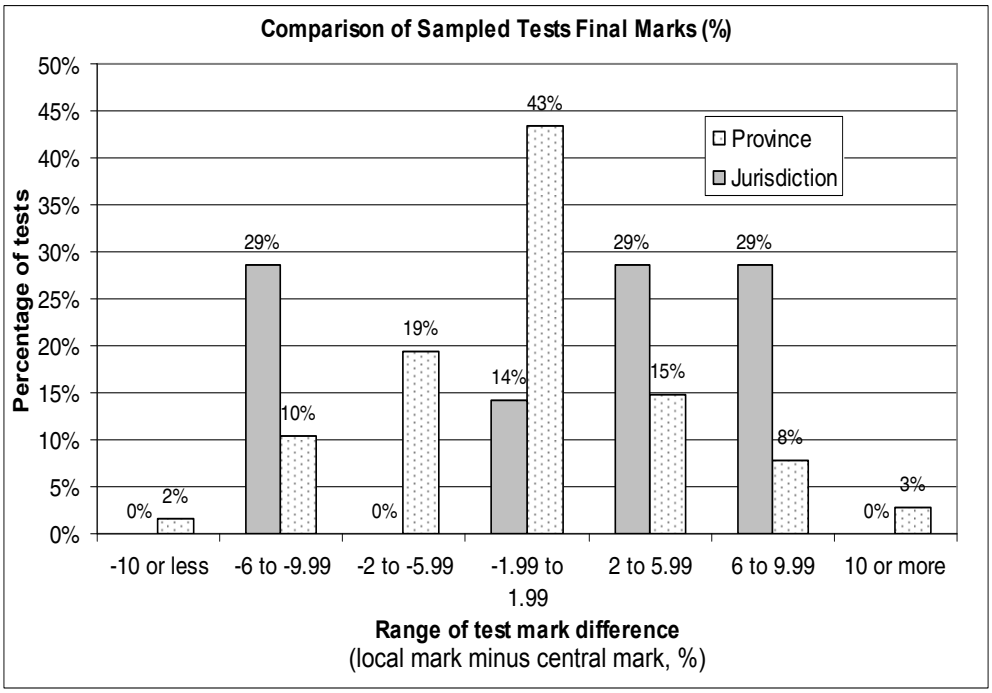
Jurisdiction: $n = 0$ $n = 1$ $n = 0$ $n = 2$ $n = 3$ $n = 1$ $n = 0$

Across the province 43% of the sampled test booklets were given the same marks locally and centrally, to within +/- 1.99 percentage points, while for 15% of the booklets, the local test marks exceeded the marks given centrally by between 2 and 5.99 percentage points. There was no consistent bias towards giving either higher or lower marks locally, evident from the bell-shaped pattern of the bars for the province (tall in the middle, decreasing in height on each side), with the longest bar (43%) at the smallest test mark difference (between -1.99% and 1.99%).

For the jurisdiction (grey bars), local markers tended to be generous. This is evident from the fact that the tallest grey bars are at or to the right of the centre with 57% (43% + 14%) of the sampled test booklets receiving higher test marks locally than centrally. Meanwhile, the bell-shaped pattern of the grey bars indicates a degree of marking *consistency*. The *n*-values beneath the chart indicate the number of sampled test booklets falling into each test mark difference category for the jurisdiction.

Following is an example of inconsistent marking in the jurisdiction (grey bars).

Example 2: Marking Consistency: Test Scores



Jurisdiction: $n = 0$ $n = 2$ $n = 0$ $n = 1$ $n = 2$ $n = 2$ $n = 0$

For the jurisdiction, local marking for the sampled test booklets (grey bars) was inconsistent. This is evident from the lack of a bell-shaped pattern. Some booklets were marked stringently (29% at -6% to -9.99%) and some were marked generously. Probable causes include a range of interpretations of marking criteria, inattention to the criteria, or markers being affected (biased) by factors not associated with the quality of students' work.

Generally, the sample sizes used in these analyses are small, and so caution is required in making interpretations. As with all sample-based analyses, there is the possibility that, by chance, the sample does not accurately reflect marking in general in your jurisdiction. It is best practice to consider the results in this report in conjunction with factors pertaining to marking in your jurisdiction and to previous marking feedback results.

Appendix 3: Dealing with Statistical Outliers

An outlier is a data point (test score, for example) that stands out from the general pattern or distribution of the rest of the data. Generally, an outlier is easily explained. If it is a data-gathering error, then the error can be corrected. Looking for outliers is common practice for verifying that data is correct prior to any analyses.

If the outlier is an accurate data point, then eliminating it may not be appropriate. Meanwhile, its presence may affect the results of analyses leading to incorrect interpretations and decisions. This might be the case, for example, for a student who performed unusually poorly on one test—such a result may affect the student's final grade, and the final grade would not accurately reflect the student's overall knowledge and skill if very few scores are used to arrive at it.

One method of handling outliers when looking at average scores (for a group or for one individual) is using the **median** rather than the **mean** (or average). The median, like the mean, is a measure of central tendency of a group of scores, but it is determined differently than the average. The median is the middle score in a list of scores ordered from lowest to highest.

The mean and median are illustrated below using a list of seven test scores, all expressed in percent and ordered from lowest to highest for convenience:

20%, 65%, 70%, 75%, 80%, 80%, 85%

mean = 67.9%

median = 75% (the middle score in the list)

(For an even number of scores, the median is the average of the middle two scores.)

Assuming these scores are the summative test scores for a student, the median is potentially a more accurate representation of this student's overall performance. The median is one tool that can be used to manage outliers in data.

Principles and procedures for summarizing and reporting student grades are handled thoroughly and clearly in Ken O'Connor's *The Mindful School: How to Grade for Learning*.

References

Manitoba Education and Training. *Senior 4 English Language Arts: A Foundation for Implementation*. Winnipeg, MB: Manitoba Education and Training, 2000.

- - -. *Success for All Learners: A Handbook on Differentiated Instruction: A Resource for Kindergarten to Senior 4 Schools*. Winnipeg, MB: Manitoba Education and Training, 1996.

Manitoba Education and Youth. *Language of Planning in Education—Draft*. 2003. <www.edu.gov.mb.ca/k12/agenda/docs/glossary.pdf>. 6 Sept. 2006.

Manitoba Education, Citizenship and Youth. *Grade 3 Assessment in Reading, "Lecture" and Numeracy and Grade 4 Assessment in French Immersion "Lecture" / Évaluation de « Reading », de la lecture et des notions de calcul des élèves de 3e année et évaluation de la lecture des élèves de 4e année d'immersion française.* Winnipeg, MB: Manitoba Education, Citizenship and Youth, 2004. Available online at: <www.edu.gov.mb.ca/k12/assess/publications.html>.

- - -. *Policies and Procedures for Standards Tests*. Winnipeg, MB: Manitoba Education, Citizenship and Youth, Annual publication. Available online at: <www.edu.gov.mb.ca/k12/assess/publications.html>.

- - -. *Rethinking Classroom Assessment with Purpose in Mind: Assessment for Learning, Assessment as Learning, Assessment of Learning*. Winnipeg, MB: Manitoba Education, Citizenship and Youth, 2006. Available online at: <www.edu.gov.mb.ca/k12/assess/wncp/index.html>.

O'Connor, Ken. *The Mindful School: How to Grade for Learning*. Arlington Heights, IL: Skylight Training and Publishing Inc., 1999.

Printed in Canada
Imprimé au Canada