REPORT FROM THE FIELD

SETTING STANDARDS FOR A PROVINCIAL LITERACY ASSESSMENT IN SASKATCHEWAN: PREMISES AND PROCEDURES

DARRYL HUNTER *Saskatchewan Education*
TREVOR GAMBELL *University of Saskatchewan*

ABSTRACT. As pressures for public accountability mount, and large scale assessment programs multiply across Canada, questions arise about who and how educational standards may be determined assume importance. National and provincial standards are now being or are about to be defined. This article discusses the assumptions and considerations involved when choosing a standards-setting model for the Saskatchewan provincial assessment program, describes the actual procedure employed, and identifies some of the technical, legal, and reporting issues involved in standards-setting.

RÉSUMÉ. À mesure que les pressions de responsabilisation publique augmentent et que les programmes d'évaluation à grande échelle se multiplient au Canada, on commence à élaborer des questions sur qui établit les attentes éducatives et comment. Les auteurs de cet article analysent les hypothèses et les paramètres qui entrent en jeu dans le choix d'un modèle d'établissement des attentes pour le programme d'évaluation de la Saskatchewan; ils décrivent les modalités utilisées et précise certains des problèmes techniques et juridiques que pose l'établissement des attentes.

Defining educational standards in "basic skill" areas is an emerging issue Canada-wide. Many provinces are turning to large-scale assessment programs as a crucible for determining formal and publicly proclaimed standards of student achievement, and as a source of information about the performance of educational systems. Both provincial and interprovincial assessment programs are proliferating. Manitoba and Ontario have recently announced ambitious testing programs to be administered in 1996 and many other provincial ministries have instituted large-scale testing or are making plans to introduce new programs. Interprovincially, the four Atlantic provinces have decided to administer cross-jurisdictional tests. And the Council of Ministers of Education, Canada's national School Achievement Indicators Program

(SAIP), has undertaken to test 13- and 16-year-old students across the country, beginning with mathematics in 1993, reading and writing in 1994, and science in 1996. To clarify pan-Canadian expectations for SAIP results, a standards-setting process is under active consideration.

For all large-scale assessments, the definition of performance standards is a central evaluative activity. If performance levels set out the basic units for measuring the quality of student work, performance standards are judgments about expected student achievement. Rather than being a description of what students are expected to be able to do, standards are pronouncements on how well students should be able to perform. Standards provide a comparator for educators and for the public against which to measure student performance. If only test results were reported, the question would remain, "Should we be satisfied with the results?"

Answering that question begs others. How does one arrive at standards? Should the standards be set by teachers as professionals, by ministry officials delegated as those responsible for provincial programs, or by members of the community at large? How does one communicate standards to which students will aspire, that which teachers can readily apply in their classrooms, and in which the public may have confidence? Should national standards precede, succeed, or, in some cases, supersede provincial standards-setting activities?

Since standards-setting is now attracting the attention of national policy makers, but is little understood outside the realm of measurement specialists, it may be useful to consider how these issues have thus far been addressed in one Canadian setting. This article discusses the assumptions and considerations involved when choosing a standards-setting model for a Saskatchewan provincial assessment program, describes the actual procedure employed, and identifies some of the technical, legal, and reporting issues involved in standards-setting.

Assumptions about performance standards for the PLAP

Saskatchewan's Provincial Learning Assessment Program (PLAP) was designed to monitor student outcomes like other large-scale assessment programs. Initiated in 1993 and implemented in language arts in 1994, the program seeks to determine student achievement levels, to provide empirical data which will assist with program improvement, to support educators in evaluating students, and to demonstrate the Department's commitment to public accountability. More specifically, the 1994 as-

sessment in reading and writing collected baseline provincial data for tracking mainstream student progress toward the goals and linguistic objectives of provincial curricula. It also provided a diagnostic picture of student strengths and weaknesses in both basic and higher order skills, and set performance standards or expectations for language arts achievement.

As conceived in the Saskatchewan program, a standard is a dynamic tool to facilitate educational decision-making. The cut-off scores or thresholds are conceived as points of educational decision-making for the next two years. When the 1996 PLA is conducted, standards will be set again and biennially thereafter for future cycles of the assessment program. Thus, standards-setting involves setting expectations for student performance in reading and writing, for a specific battery of tests or assessment instruments. Standards or expectations are not static. Rather, expectations will vary according to the complexity of the questions or tasks demanded in various tests, to the criteria used for scoring the tests, the attributes of the student population being tested, and to the composition of the panel which sets the standards. Because the PLA is a criterion-referenced assessment and not a norm-referenced one, a provincial average is an unsuitable yardstick. A criterion-referenced standards-setting model was therefore devised, based on a number of premises.

A primary assumption for the project was that an existing standards-setting procedure, extensively employed in other settings and test situations, could be adapted to set valid standards for the Saskatchewan project. This inevitably led to an examination of the literature from the United States, where minimal competency standards have been established for many state-wide criterion-referenced assessments. Ultimately, the single cut-off score as a standard associated with minimal competence was rejected. This type of standard fails to recognize a range of student ability, tends to undermine rather than reinforce curricula, and has the effect of disadvantaging children from impoverished and minority families when used in "high-stakes" decisions (Jaeger, 1989). Having a standard which provides multiple decision-making points or cut-offs was deemed more useful useful to teachers, administrators, and policy makers who make program decisions to meet a variety of student needs.

Second, the standard was to remain closely bound to the context in which it was set, and the collection of assessment instruments from which it was developed. The standard was to be derived by looking at

the specific pool of items used on tests, the Table of Specifications for the assessment, the scoring criteria and procedures employed, and the specific characteristics of Saskatchewan's population at the time the standard was set. Thus, the standards set for the 1994 PLA must not be viewed as "provincial" standards in reading and writing. They are the PLA standards for grades 5, 8, and 11 – the grade levels at which the assessment was administered in 1994. They were set in relation to the current pool of items, tasks, exemplars, and criteria used to determine knowledge, skills, and attitudes specified in 1993. Nevertheless, insofar as the 1994 PLA reflects a provincial consensus on those literacy behaviours and skills that students should possess, and until such time as the Department formally incorporates the performance levels in curricula, the levels will likely be perceived as functional standards for the provincial educational system.

This leads to a third and important assumption: although communication of the standards may not ultimately lead to improved provincial performance, students will respond to clearly delineated expectations, with detailed descriptions of criteria which are in turn linked to examples of challenging reading and writing tasks. At the same time, teachers will modify their instruction to conform to provincial curricular objectives in fostering these skills. Moreover, administrators will mobilize school resources to those ends, if the standards clearly articulate what is expected of students. The performance levels, in conjunction with the expectations assigned each, were thus designed not only to respond to questions of public accountability, but also to resonate in classrooms.

In this way, the standards would tangentially serve as a target to ultimately accelerate curriculum implementation. In Saskatchewan, assessment and evaluation are seen as integral to – not as appendices on, nor antithetical to – curriculum and instruction. If one designed an assessment project which was based on the foundational learning objectives in current and impending curricula, and if performance criteria were linked to these objectives, then the standards would reinforce curricular and instructional goals. As such, the standards would serve both as a fulcrum for moving the field toward provincially defined educational aims and goals, and as a means of enabling the province to calibrate its scoring and grading practices.

A fifth assumption follows closely on the above: even though the standards would encourage schools and school divisions to adopt provincial curricular and instructional goals, standardization would not

result. Instead of reporting only average performance, or minimally acceptable performance, the PLA would identify different standards or expectations for different performance levels. A variety of new and useful interpretations would be possible to facilitate a broader variety of educational decision-making. For example, test results indicated that only 2% of students reached level 5, top proficiency in organizing their writing, where the writing demonstrated a purposeful and effective order and arrangement of events, ideas, and details, whereas 11% were expected to attain this level. On the other hand, 6% of students achieved level 1 where the writing demonstrated an unclear or haphazard order and arrangement of ideas, events, and details, "exceeding" expectations of 5%. The message sent to students, parents, teachers, administrators, and policy makers would be clear: when only 2% of grade 11 students reach level 5 proficiency in their organizational skill in writing, considerable attention must be paid to improving this skill. While only 6% of students reach level 1, this is within the range of expectations; we will never be able to have the entire population reach a mastery level because of the circumstances in which youth are sometimes placed.

Indeed, the PLA did not assume that standards are synonymous with educational quality, an equation that is often made in the rhetoric of popular discourse. Elevating test standards, so the argument goes, will improve the quality of educational instruction. However, as the Quebec Ministry of Education discovered in 1986-1987 after raising the passing scores on its provincial examinations from 50% to 60%, one of the principal effects was to increase the drop-out rate from its schools (Maheu, 1995, p. 60). Raising the threshold for defining competence does not result in better educational outcomes for all.

Who has responsibility for determining standards?

Selection of judges has received insufficient attention in the standards-setting literature. Yet as many studies suggest, the validity of standards is directly related to the qualifications of the judges. One study states that judges should be drawn from those who possess expertise in the subject domain being assessed (Jaeger, 1991); others suggest they should be selected from those constituencies which have a stake in the application of standards (Hambleton & Powell, 1983). These different selection criteria are not always compatible (Bourque & Hambleton, 1993).

Because standards would be generalized across the province, and because Saskatchewan has a long tradition of collaboration among the

educational stakeholders, the Department decided that standards-setting activity should involve both educators and noneducators. Accordingly, letters of invitation were sent to the executives of the various stakeholder organizations, requesting delegates. General qualifications were a collaborative approach to decision-making, general experience in working with youth, and acceptance of purposes of the PLA program.

To set standards, a 26-member standards committee was assembled, consisting of representatives of the major education stakeholders who worked in three grade-level subcommittees. Balance was sought when designing the exact composition of each standards subcommittee, so that the various gender, geographic, ethnic, and stakeholder groups were represented, and so that their collective expectations might reflect those of the various constituencies in the provincial educational system. Each subcommittee included:

- Three teachers chosen to represent urban, rural, and northern school situations;
- One representative from the Saskatchewan Department of Education, Training and Employment (Curriculum and Instruction Branch);
- One representative of the Saskatchewan Teachers' Federation, customarily a teacher;
- One representative of the Saskatchewan School Trustees' Association, customarily a parent-trustee;
- One representative of the League of Educational Administrators, Directors, and Superintendents;
- One representative of a postsecondary institution such as the University of Regina or the University of Saskatchewan, and the Indian-Métis Education Advisory Committee; and
- One representative of the business community.

Aboriginal representation was included within each subcommittee's nine-member structure.

If there were initial concerns that the non-educators delegated by stakeholders to the standards panel would not have expertise in language arts and its issues, this apprehension was soon allayed. Those nominated were invariably former educators, were non-educator representatives on other curriculum reference groups, or were knowledgeable about language arts curriculum issues through work on school boards.

Thus, stakeholders selected representatives who had some experience or expertise in the subject area.

Choosing a standards-setting procedure: Rationale and revision

A number of criterion-referenced standards-setting procedures have been developed for use in many test situations. Berk's exhaustive 1986 survey of the field identified some 38 different methods. However, for the PLA program, the number of available options was restricted because of the characteristics of the assessment design. Some methods like the "contrasting groups method" or "the borderline groups method" (Livingston & Zieky, 1982) require identifying specific groups of students or identifying in some other way the academic performance of individual students. Because the assessment's purpose was to create a provincial profile of achievement through random sampling, and because the anonymity of participating students, teachers, schools, and school divisions was ensured, individual student information was not available for these methodologies to be employed.

Three other methods were ruled out because of the variety of items employed in the provincial assessment. Ebel's method (1972) involves asking judges to make two types of decisions about test items: 1) a judgment of each question's difficulty and 2) a judgment of its relevance or importance. Using these two different scales as axes on a matrix, judges then classify and group questions into the matrix cells, before estimating the percentage of questions that competent students would answer. On the other hand, Jaeger's method (1978) simply asks judges to consider each test item with a question about whether students should or should not be able to answer it. Nedelsky's method (1954) asks panelists to make judgments about potential response to multiple-choice questions. All three methodologies were unsuitable for the PLA program which uses extended open-response essays, multiple-choice items, and short-answer items; none of these methods could accommodate open-response questions.

The selected method, therefore, had to be flexible enough to accommodate a variety of item types, be adaptable for the task of setting multiple cut-off points so as to provide appropriate classification information for decision-makers, be easy to compute, be credible to lay people, be recognized in measurement literature as statistically sound, and be relatively straight forward in implementation. Two major reviews of standards-setting methods (Berk, 1986; Jaeger, 1989) both recommended the Angoff (1971) method as simultaneously meeting these require-

ments, while having the added credibility of being the most extensively used method in American testing programs.

When adapting Angoff's method for the Saskatchewan setting, however, three important changes were made. The first involved modifying the question. In his original formulation, Angoff asked judges to estimate the probability that a competent test-taker would answer the question correctly. Because of the apprehension that non-educator judges would not have sufficient direct experience with youth to make realistic estimates of student capability, and because the Department wanted a projection of the potential for Saskatchewan youth, the question was modified to use the word "should" instead of "would". Using this wording would define standards for desired performance, rather than simply identifying thresholds of actual student competence.

A second adaptation involved anchoring judges' decisions directly in examples of student work, rather than only in the test items, tasks, and performance criteria. Exemplars of student performances which illustrated each point in the scale were presented to judges, along with the criteria, so as to give both educator and non-educator judges a clear conception of the range of student ability and how the actual scoring criteria had been operationalized in the preceding scoring session. In this sense, the learning assessment involved two types of evaluative activity: the scoring judgments made by teacher-specialists in categorizing student performances and the standards-setting panel's subsequent judgment of overall student competence in the province for its strengths and weaknesses.

The third adaptation, however, highlighted a difference between these two phases of evaluative activity in the assessment program. The standards-setting exercise was designed to be a three-round iterative activity, to allow for judges to exchange opinions in a controlled fashion, to allow judges' opinions to stabilize for the sake of reliability, and to provide for an eventual consensus to emerge. Whereas the scoring session had demanded that a group consensus or interpretive community precede the assignment of mathematical scores according to the evaluative scale, the subsequent standards-setting session provided for the mathematical assignment of estimates to precede the emergence of a group consensus. Indeed, the standards-setting procedure anticipated quite divergent expectations to be expressed, and was thus not predicated on a *consensus ad idem* or meeting of minds. Consensus was a product of, not a prerequisite to, participation in the standards-setting process.

To merge the various estimates provided by judges, one of three mathematical methods of compromise had to be chosen: a mean, a median, or a trimmed mean. An average would enable the participating stakeholders to have a direct stake in the standard, but a mean is more susceptible to distortion by aberrant judges whose patterns of estimations might be inconsistent with others'. A median or trimmed mean, on the other hand, deliberately removes the extreme estimates and reduces the effect on the resultant cut-off scores. However, discarding the outlying estimates may be at odds with the inclusionary purposes of incorporating stakeholders in a standards-setting panel. For this reason, and because it "is well known that the sample median is a less stable statistic than the sample mean" (Jaeger, 1991, p. 14), particularly for a small sample, the average was chosen to combine the judges' various recommendations and to compute expectations for each of the five performance levels.

Standards were established for five dimensions of writing and five types of higher-order skills associated with reading, as well as for reading comprehension and writing as a whole, using the modified Angoff method. Standards were developed in three stages, with subcommittee members basing their judgments directly on actual student work and scoring criteria. Facilitators were recruited from among the scoring leaders and assessment designers to provide judges with insight into the construction of assessment items and their scoring.

First, the facilitator reviewed the reading or writing task, described the criteria used in scoring the item(s), and presented examples of student work at each performance level. This was a blind review. Actual results for the assessment were not divulged. The facilitator then asked, "In this skill area, which percentage of the regular stream school population should be able to attain each performance level?" Without consulting others, each member privately wrote down on a tally sheet his or her preliminary estimate of proportions of students who should have been able to attain each of the five levels. These estimates were collected, and a mean distribution was calculated and distributed to all group members.

In the second stage, judges were in turn invited individually to provide comments on the preliminary mean distribution. Comments were restricted to the nature or complexity of the task or questions, the criteria used for scoring student work, the examples of student work presented, and the attributes of the school population being tested. Once every

judge had spoken, a short discussion was conducted to allow additional viewpoints to be expressed. Members were then given the opportunity to revise their preliminary estimates in light of the insights and comments generated by the panel. The revised estimates were written down on a tally sheet, collected and averaged to produce a revised mean distribution.

The third stage was an informed review. Judges were given actual student results along with the revised mean distribution they had provisionally set as a standard. Each member was again invited in turn to provide comments on the committee's revised mean distribution. Having heard everyone speak, each member was allowed a second opportunity to revise privately his or her expectations in light of the comments made and the actual results presented. Tally sheets were collected and a mean distribution of the group's individual expectations was calculated to generate provincial standards for the reading and writing skill area under consideration.

Issues in setting standards for the provincial LAP.

When embarking on the provincial assessment program, new language arts courses were under development. A new curriculum had been developed and implemented in Saskatchewan schools at the elementary level, but was still on the drawing board for the middle years and high school levels when the assessment was designed in 1993. The Table of Specifications for the assessment had to strike a compromise between current curricula and blueprints for curriculum reform. This delicate balance between what is and what will be was reflected in the assessment design.

If the absence of expertise proved an issue for non-educator judges in setting standards for the 1990 American National Assessment of Educational Progress in Mathematics (Bourque & Hambleton, 1993), it was rather the level of expertise of non-educators in language arts curricular issues which was the source of some difficulty for a few judges in the Saskatchewan Provincial Language Arts Learning Assessment exercise. Nominees came to the table with questions about emphases in upcoming curricula, with questions about instructional practices, and with general questions about educational equity. These questions needed to be addressed. When these issues emerged in panel discussions during the exercise, some panel members were distracted from the main task at hand, that is, adjudicating student performance on the 1994 Provincial Learning Assessment in Language Arts.

The tension inherent in assessment design between curricular reality and curricular intent was also expressed in discussions around the question to which judges were asked to respond: "Which percentage of the regular stream Saskatchewan student population should be able to attain each of the five levels?" If the question of what students "should be able to do" asks for an estimation of student potential or description of idealistic expectations, the question of what students "would be able to do" describes student competence in light of current classroom realities. Considerable discussion revolved around the meaning of the term "should"; is it viable to set standards which describe student performance in optimal circumstances, or would it be better to define thresholds of acceptability given the educational system's current status? The former question may be best for the purposes of establishing baseline data and program enhancement, whereas the latter question may be more appropriate for public accountability purposes.

For a very few judges, the procedure itself, with its focus on producing percentage distributions, was artificial. Underlying their concerns was the perceived subjectivity of the judgments rendered, despite the performance data and examples of student work provided. Popham (1978) has described this oft-voiced concern succinctly: "Unable to avoid reliance on human judgment as the chief ingredient in standards-setting, some individuals have thrown up their hands in dismay and cast aside all efforts in setting standards as arbitrary, and hence unacceptable" (p. 168).

But Popham goes on to identify the *non sequitur* in this reasoning by providing two dictionary definitions of "arbitrary": "The first of these is positive, describing arbitrary as an adjective reflecting choice or discretion, that is determinable by a judge or tribunal. The second definition, pejorative in nature, describes arbitrary as an adjective denoting capriciousness, that is, selected at random without reason" (p. 168). In Popham's opinion, "when people start knocking the standards-setting game as arbitrary, they are clearly employing Webster's second, negatively loaded definition. But the first definition is more accurately reflective of serious standards-setting efforts. . . That they are judgmental is inescapable. But to malign all judgmental operations as capricious is absurd" (p. 168).

Standards may not be objectively determined, but they can be objectively applied. Indeed, one of the primary purposes of setting standards is to provide for fairness and impartiality when making educational

decisions. That equitable application has been the heart of landmark legal rulings in the United States. The series of four court cases which comprise *Debra P. versus Turlington* were decided between 1979 and 1982 with reference to the Fourteenth Amendment guarantees of due process and equal protection. These state that: "No State shall make or enforce any law which shall abridge the privileges or immunities of the United States; nor shall any State deprive any person of life, liberty or property, without due process of law; nor deny to any person within its jurisdiction the equal protection of laws". These constitutional provisions have been held by the American courts to provide protection to students for whom competency tests are used to make educational decisions. The American Fifth Circuit Court held that a state "may not constitutionally so deprive its students [of a high school diploma based on test performance] unless it has submitted proof of the curricular validity of the test". The Court further determined that "if the test covers material not taught the students, it is unfair and violates the Equal Protection and Due Process Clauses of the United States Constitution" (Jaeger, 1989; Gunn, 1982; Logar, 1984).

Canadian courts may turn to American judicial reasoning for guidance on these matters, as they have done in other recent school-related cases, because the Charter of Rights and Freedoms contains two provisions which are similar to American constitutional guarantees. Section 7 of the Charter states that: "Everyone has the right to life, liberty and security of the person and the right not to be deprived thereof except in accordance with the principles of fundamental justice". And Section 15(1) recognizes that: "Every individual is equal before and under the law and has the right to equal protection and equal benefit of the law without discrimination and, in particular, without discrimination based on race, national or ethnic origin, colour, religion, sex, age or mental and physical disability". In particular, the clauses setting out "principles of fundamental justice" and "equal benefit of the law without discrimination" suggest that educators may be subject to the same legal tests that have been used for defining the educational standards in American competency tests. Courts on both sides of the border have an historic posture of judicial restraint in addressing school-related issues, viewing them as the province of school officials. Yet the Charter may draw into question students' opportunity to learn the tested material, a test's differential impact with gender and ethnic groups, and the principles of fundamental justice when determining and applying standards (Hunter & Matthews, forthcoming).

Thus, equity issues are a central feature of any standards-setting activity, and the legal implications are important. Even though the provincial assessment was a "low-stakes" assessment involving random samples of students from which generalizations could be made about provincial literacy performance, the standards may be extrapolated to individual schools and school divisions for such "high-stakes" decisions as promotion. Legal sanctions might not apply to the original standards-setting activity, but become paramount depending on the use to which the standard is subsequently made. These questions place an added responsibility on those who undertake to define performance standards to ensure that the method is technically sound, and allows for due process.

How judges arrive at their decisions

An element of due process is reasoned choice. A standards-setting method can be conceived as a formal mechanism for supplying judges with evidence on which they base their evaluations of student performance. The effectiveness of any procedure may depend on how efficiently it supplies the types of information setters need to make decisions reliably. Some procedures have been criticized because of their subjectivity; judges may have the sense that they are "pulling their probabilities from thin air" (Berk, 1986, p. 147), rather than making reasoned choices based on evidence. No studies have been published which examine the *ratio decidendi* of judges who have set a criterion-referenced standard.

At three points in the Saskatchewan process — after orientation but before the first stage of balloting, after the second stage of balloting, and after the third and final stage of exercise — judges were asked to rank order ten different types of information provided during the exercise as most and least useful in their decision-making. These types of information included:

- Descriptions of scoring procedures and criteria.
- The test questions and tasks.
- The standards or expectations expressed by the organization or institution the judge represented.
- Initial opinions and viewpoints expressed by other subcommittee members.

- Direct professional experience with youth including contact in the classroom, the work-force, or the personnel office.
- Group discussions which followed initial comments made by subcommittee members.
- Direct personal experience with youth, including assessment of their own children's abilities.
- Actual student achievement results.
- Statistical information, including indexes of difficulty for multiple-choice items.
- Examples of student work at each of the five performance levels.

Although the design of the standards-setting exercise did not allow individual judges' evidentiary bases to be partitioned out, broad patterns are evident when examining judges' overall ratings of the three most useful and three least useful types of evidence compared across the various stages of the Saskatchewan process.

Not surprisingly perhaps, the test questions and tasks, descriptions of scoring procedures and criteria, exemplars of student work, and direct professional experience were anticipated to be the most useful types of information before the actual exercise began. At the same time, approximately half of the judges ranked their sending organization's standards as among the least useful. Statistical information and direct personal experience with youth were the other two types of information most frequently anticipated as being least useful.

When surveyed after the second stage of the process, descriptions of scoring procedures, examples of student work at each of the performance levels, and direct professional experience were the most useful types of information on which judges based their estimates. Indexes of difficulty for multiple-choice questions, the anticipated student achievement results, and the standards or expectations expressed by the institution or organization each judge represented were consistently chosen, at this mid-point in the exercise, as the least useful evidence for basing probability estimates.

However, after the exercise, judges most frequently cited the test questions and tasks, and their direct personal and professional experience with youth as being the most useful in their decision-making. Deemed by judges before the session as being among the least useful, personal

experience with youth became among the most useful bases for making probability estimates by the end of the session. Statistical information, descriptions of scoring procedures, and the standards expressed by their sending organization were viewed as least useful. The examples of student work and the descriptions of scoring procedures and criteria had receded in perceived usefulness. Only seven of the 26 judges reported, after the exercise, that the actual student results were among the top three types of information they found useful in making their decisions.

What conclusions can we draw from these shifting viewpoints on evidentiary usefulness? First, the opinions of those incorporated into a standards panel did not seem to be shaped by the organizational agendas of stakeholders, either before, during, or after a standards-setting session. Second, while professional expertise with youth is frequently a basis for decision-making at all stages of the process, direct personal experience with youth, including assessment of one's own children and their capabilities, increasingly became a useful referent when offering estimates. The fact that group discussions and the opinions of others did not fluctuate, although they are consistently considered in the mid-range of usefulness, suggests that group exchanges do not significantly alter judges' opinions. But third, and most important, the statistical information, either in the form of item statistics or the actual student performance data, did not have a preponderate weight in shaping judges' final estimates. Taken together, the survey suggests that when making summative judgments of what students should be able to do, in describing student potential, judges will rely more on their personal and professional experience in light of the test questions and tasks, rather than on statistical evidence, evidence of student work, or actual student performance data that have been presented during a standards-setting session.

When asked to base their opinions in evidence, judges frequently and increasingly anchored them in previous experience rather than in information supplied directly by the standards-setting procedure. If that is so, the procedure used in Saskatchewan may have served to solidify individual preconceptions of student capability, rather than fundamentally change them in light of actual data compiled by the assessment. In other words, the controlled exchange of viewpoints and the careful presentation of documentary and statistical performance may have affirmed rather than substantially altered judges' pre-existing patterns of reasoning.

How standards should be reported

In addition to unanswered questions about selecting judges before a standards exercise, and the reasoning processes of judges during an exercise, many issues remain about reporting standards after an exercise. The graphic display chosen can have an important impact in communicating results and expectations. For the Saskatchewan reports, two alternate displays were chosen to communicate results to two different audiences in a manner compatible with the varying purposes for the assessment.

TABLE 1. Grade 11 writing performance – overall

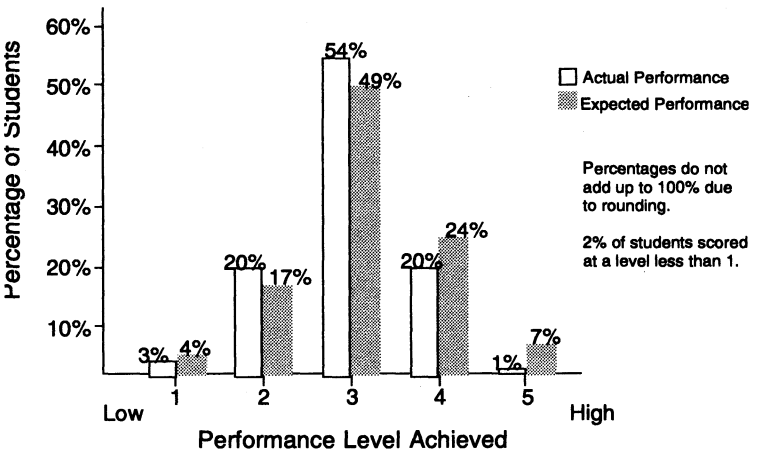
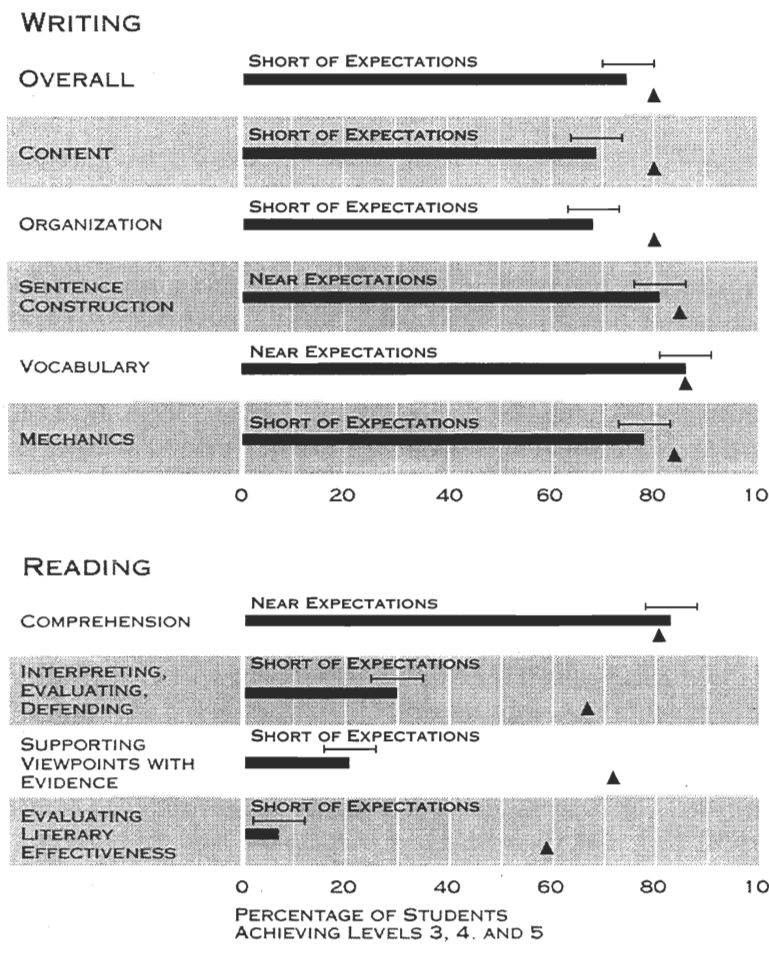


Table 1 illustrates the graphic presentation chosen for reporting results in a document destined for the educators. By presenting the individual expectations for each of the different performance levels, the table was designed to reflect closely the actual judgmental process employed in the standards-setting exercise. The intent was to covey the message that there is a range of acceptable performances consistent with the range of student abilities in a typical classroom. The actual performance data approximate the normal distribution of student ability while simultaneously suggesting that there is potential for growth in the upper performance levels. By stating expectations for each performance level, administrators and educators are given a variety of decision-making points for classroom groupings or streaming students into different programs. References to criteria for each of these performance levels, with examples of student work, would further serve to categorize and operationalize student achievement.

TABLE 2. Grade 11 language arts achievement in brief, 1994



Note - The margin of statistical error represents the percentage by which the results of the survey can be inaccurate.

Source: Saskatchewan Provincial Language Arts Learning Assessment. Anticipated publication, Fall 1995.

While Table 1 may be most appropriate for program enhancement and monitoring student growth, Table 2 was explicitly designed for purposes of public accountability. As a condensation of grade level results for inclusion in the Saskatchewan Education Indicators Report, the graphic display was designed to be one outcome indicator of systemic performance. Modeled after that used for the Council of Ministers of Education, Canada's reports on the 1994 School Achievement Indicators Program, the display conveys in a more linear fashion the strengths and weaknesses of student performance in a range of literacy skills. At the same time, it illustrates the margin of error and the inherent imprecision of any large-scale assessment, which samples on a random basis, in order to qualify the sometimes uninformed opinions of non-educators about test results. Because the term "benchmark" has entered the popular lexicon of student achievement, a triangle was chosen to indicate the combined expectations for the top three performance levels. Implicit, then, in this aggregation is the notion that student performances at levels 1 and 2 are unacceptable, an assumption that did not underlie the original assessment design but which was a feature of the standards committee's discussions and decisions.

Conclusions

Although British Columbia and Alberta have started to work in the area of standards-setting for criterion-referenced assessments, it is, for the most part, a pioneering activity in Saskatchewan and other provincial settings. While other jurisdictions will likely turn to models that have been devised in the United States with minimal competency legislation, significant adaptations will need to be made when transplanting them into Canada. A host of technical, legal, and reporting questions will have to be systematically addressed because the purposes for assessment programs, and the educational environment in which they will be conducted, are quite different than those in the United States.

Saskatchewan's experience with this pivotal evaluative activity suggests that stakeholders, including non-educators, can be meaningfully incorporated in a standards panel. The question which prompts these panelists in their standards-setting activity requires careful consideration, since the resultant standard may hinge on its exact wording. So too must the evidence on which both educator and non-educator judges make their decisions be carefully examined to determine whether the chosen procedure has abetted or inhibited reasoned choice (Kane, 1994). Neither this weighing of evidence nor the legal dimensions of

standards-setting have been sufficiently explored. To determine the reliability of judges' decision making, psychometricians have thus far concentrated on statistical analyses of judges' voting patterns, through standard deviations and indices of variability considering this as a measure of a method's trustworthiness. However, a method's effectiveness can also be viewed in terms of its efficacy in supplying useful data to participants. And finally, while the preparatory and procedural elements of designing an exercise are important, how one packages and communicates the product of an exercise also deserves systematic consideration.

All educators recognize that the product is an important lever of public policy and educational change. If national education standards are to be defined, the constitutional basis of standards-setting activities must be determined; the rights extended by the Charter of Rights and Freedoms and the extensive litigation in the United States, regarding the student's opportunity to learn material on which he or she is being tested, may be considerations if provincial ministries choose to use standards for high stakes purposes. Ensuring that the standards effect positive changes in curriculum and instruction for students, rather than distorting the educational agenda, will be the challenge for both policy makers and educators in the provincial and national arenas.

ACKNOWLEDGEMENT

This project was funded by the Social Sciences and Humanities Research Council of Canada #410-91-0050. The support provided by the teachers from the four divisions sampled for the study was also greatly appreciated.

REFERENCES

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thomdike (Ed.), *Educational measurement* (2nd ed. pp. 508-600). Washington, DC: American Council on Education.

Berk. R.A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56*(1), 137-172.

Bourque, M.L., & Hambleton, R.K. (1993). Setting performance standards on the national assessment of educational progress. *Measurement and Evaluation in Counselling and Development, 4*(26), 41-47.

Ebel, R.L. (1972). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Gunn, L.D. (1982). Debra P. v. Turlington: Due process enters the classroom, but how far? *Journal of Law and Education, 11*(4), 573-585.

Hambleton, R.K., & Powell, S. (1983). A framework for viewing the process of standard-setting. *Evaluation and the Health Professions*, 6(1), 3-24.

Jaeger, R.M. (1978). *A proposal for setting a standard on the North Carolina High School Competency Test.* Paper presented at the meeting of the North Carolina Association for Research in Education.

Jaeger, R.M. (1989). The certification of student competence. In R.L. Linn (Ed.), *Educational measurement* (pp. 485-513). London: Collier Macmillan.

Jaeger, R.M. (1991). Selection of judges for standard setting. *Educational Measurement: Issues & Practices, 10*(2), 3-14.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425-461.

Livingston, S.A., & Zieky, M.J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Logar, A. (1984). Minimal competency testing in schools: Legislative action and judicial review. *Journal of Law and Education, 13*(1), 25-49.

Maheu, R. (1995). Education indicators in Quebec. *Canadian Journal of Education, 20*(1), 56-64.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14*(3), 3-19.

Popham, W.J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.

DARRYL HUNTER is Director of Assessment and Evaluation, Saskatchewan Education, Training and Employment.

TREVOR GAMBELL is Professor and Graduate Chair, Department of Curriculum Studies, College of Education, University of Saskatchewan.

DARRYL HUNTER est directeur des mesures d'évaluation au ministère de l'éducation, de la Formation et de l'Emploi de la Saskatchewan.

TREVOR GAMBELL est professeur et directeur des études supérieures au département d'études sur les programmes d'études, Collège des sciences de l'éduuation, Université de la Saskatchewan.