The Canadian Journal of Program Evaluation Vol. 11 No. 1 Pages 35–61
ISSN 0834-1516

SETTING ACHIEVEMENT STANDARDS/EXPECTATIONS FOR LARGE-SCALE STUDENT ASSESSMENTS

Richard M. Jones
Ontario Education Quality and Accountability Office
Toronto, Ontario

Darryl M. Hunter
Director, Assessment and Evaluation Unit
Saskatchewan Education
Regina, Saskatchewan

Abstract: This article discusses a variety of issues associated with establishing formal standards or expectations for student achievement on large-scale assessment programs. The article reviews the educational literature on the subject of standard-setting methods, describes two approaches to standard setting that have been used in province-wide assessments in Saskatchewan, and evaluates the effectiveness of these approaches. Although the two procedures described were generally successful, numerous issues are raised and recommendations are made to support future standard-setting activity and research.

Résumé: L'article aborde une variété de questions liées a l'établissement de normes ou d'attentes précises pour le rendement scolaire dans le cadre de programmes d'évaluation à grande échelle. L'article fait une revue de la littérature éducationnelle sur la question des méthodes d'établissement de normes, décrit deux stratégies d'établissement de normes utilisées lors d'évaluations à l'échelle provinciale en Saskatchewan et évalue l'efficacité de ces stratégies. Bien que les deux stratégies décrites aient été une réussite, elles soulevent un bon nombre de questions; l'article formule des recommandations pour appuyer la recherche et les activités futures en matière d'établissement de normes.

Establishing formal standards or expectations for student achievement in large-scale assessment programs is a relatively new activity in Canadian educational circles. Once the exclusive domain of psychometricians and education measurement specialists, policy makers and educational professionals are increasingly interested in how to define acceptable and/or desirable levels of performance as provincial assessment programs are now proliferating.

A distinction must be drawn between performance levels that describe levels of attainment or what students can do, and standards that describe how many students could be expected to attain each level. Without standards to compare with actual results, both educators and the public have difficulty interpreting student performance in a meaningful fashion. If only test results and scales are reported, the question remains, "Is student performance acceptable?" In many projects, including recent assessment rounds of the School Achievement Indicators Program (SAIP), co-ordinated by the Council of Ministers of Education, Canada (CMEC), assessors have reverted to national or provincial norms as yardsticks for judging student performance, even though the assessments were criterion-referenced in their original design. While some provinces such as Alberta and British Columbia have initiated formal standard-setting activities with their provincial assessment programs, many in the educational community remain uninformed about this crucial educational activity.

Because standards are often contentious, the method must be carefully chosen and defensible if educators and the public are to have confidence in the product. This entails careful consideration of the merits of various alternatives, the questions around which the procedure is organized, the purposes for the resulting standards, the manner in which the standard will be reported, and ultimately the uses to which the standard will be employed as the point of educational decision making (Hambleton & Powell, 1983).

The purposes of this article are to review the literature on standard setting, describe standard-setting approaches associated with provincial assessments in Saskatchewan, and evaluate the effectiveness of these approaches in light of the considerations suggested in the literature.

LITERATURE REVIEW

American and Canadian Context

There is a rich vein of American educational literature on the subject of standard-setting methods, and an extended body of American case law as well. At least 37 methods were described in Berk's (1986) review of the field. Most were devised in the United States during the 1970s in concert with the introduction of minimal competency legislation, and were legally tested in American courts dur-

ing the early 1980s. Thus, there are both practical and judicial precedents to guide the measurement practitioner in selecting a standard-setting procedure.

Formal standard-setting activities in Canada, however, are still relatively rare. Few standard-setting activities have been publicly documented in the Canadian setting, and no published studies have examined their relative effectiveness north of the 49th parallel. Indeed, *Principles for Fair Student Assessment Practices for Education in Canada* stipulates only that ministries and departments of education "describe how passing and cut-off scores . . . were set and provide rates of misclassification" (Joint Advisory Committee, 1993, p. 18) when collecting and interpreting assessment information. By default, Canadian assessors searching for a standard-setting procedure will likely look outside Canada for their methods.

This is unfortunate, because most methods thus far formulated have been for the American certification of student competence. State legislation and judicial rulings have defined standard-setting methods so that they yield a single crucial cutoff score as the point of decision making about competence or incompetence, passing or failing, and certification or non-certification. In Canada, on the other hand, standards are increasingly required by policy makers to define generalizable expectations for student performance in programs and institutions, rather than for "high stakes" decisions about individual students. An American audience may find legally efficient procedures in the literature for arriving at a dichotomous decision from a single cutoff score, but the Canadian assessor will find little guidance for ascertaining multiple decision-making points derived through multiple cutoff scores.

Certainly, American measurement practitioners have chafed at the legislative strait-jacket of minimal competence. Most educators recognize that "competence is, by virtually all conceptions, a continuous variable. Setting a cut-off score that supposedly divides students into distinct categories, the competent and the incompetent, is unrealistic and illogical" (Jaeger, 1989b, p. 492). Minimal competence has distorted curricula and instruction: educators are teaching only to the level of skill demanded in the competency test and the standard established for it. More damaging have been the societal consequences. Evidence "strongly suggests that children of poor parents fail competency tests more frequently than children of rich parents. The same can be said of black children, when compared with white children" (Jaeger, 1989b, p. 510). Minimum competency testing for

high school graduation, as conducted in nearly half the states, "disproportionately affects the economic well-being of black students who aspire to attend college" (Jaeger, 1989b, p. 511). The challenge for assessors and policy-makers will be to employ standard-setting methods that recognize varying levels of ability, that reinforce rather than undermine curricula, and that promote rather than restrict the interests of the disadvantaged.

Another challenge arises from current trends in assessment techniques: most standard-setting methods thus far devised have been developed around correct/incorrect types of response items; they do not reflect the growing interest in scoring rubrics, direct writing assessment, and performance methods (Huot, 1990). As evaluators are increasingly employing holistically scored, open-response items and performance assessment techniques to evaluate student skills, they are discovering that standard-setting methods designed for the indirect, multiple-choice test formats of the 1960s and 1970s have to be substantially modified (Busch & Jaeger, 1984). Indeed, only recently have they started proposing new methods to set standards for multidimensional scaled performances (Jaeger, 1993).

The preoccupations of American assessors have shaped the standard-setting literature in other ways. Most researchers have investigated procedures for a specific testing situation, rather than placing changing standards in a larger or longer-term framework. Education indicator systems are in vogue as provincial governments and the CMEC address issues of public accountability. Changing standards plotted longitudinally may themselves become valuable long-term indicators. If judges are drawn from the same constituencies over a series of test cycles and if one employs the same method and instrumentation over those cycles, then the standards may be conceived as incarnating or embodying a set of public expectations for student performance at given points in time. Standards can thus become barometers of public- or educational-policy makers' expectations about the school system.

Even though Canadian standard-setting exercises have not been studied in the scholarly literature, pathfinding work in three provinces is described in public documents. British Columbia's Provincial Learning Assessment Program began in 1976, and has consistently employed "interpretation panels" of teachers to judge grade-level student performance in various dimensions of mathematical, scientific, and communications skill, depending on the assessment. Although the procedure has varied from one assessment

to another, constants have included the exclusive use of professionals as judges, preparatory recording of expectations as estimates for provincial performance on individual items according to "acceptable" and "desirable" categories, and subsequent formal summary and consensual judgements of performance according to four- to six-point scales ranging from unsatisfactory to excellent by the empanelled judges. (See, for example, Bateson et al., 1991, pp. 243–244; Jeroski, 1989, p. 15.)

Recent standard-setting activity in Alberta aims "to widen the process of setting assessment standards as much as possible over previous years and especially to provide for community input and feedback" (Alberta Education, 1991, p. 81). To that end, five committees have been struck as part of the Provincial Achievement Testing Program to define two standards in relation to the curriculum being tested. These committees are composed of curriculum and test developers, educational administrators, teachers from across the province, psychometricians, statisticians, and representatives from professional, business, and community organizations. Each committee is challenged to determine what score a student must obtain, or how many questions a student must answer correctly, to be judged as having achieved an acceptable and an excellent standard. A summary standard is determined by a Final Standards Review Committee, using provisional standards, review commentary, and representatives from the original five committees (Alberta Education, 1993, pp. 101–109).

While the British Columbia and Alberta ministries define standards in relation to large-scale assessment results, the Toronto Board of Education's Benchmarks Program has avoided evaluating its student population against external standards associated with a testing program. Rather, more than 100 benchmarks for language arts and mathematics have been developed as model activities for teacher emulation in the classroom setting. Based on provincial and system objectives, developed and field-tested informally by teacher committees, and emphasizing complex but observable tasks, the benchmarks set out performance levels and criteria but not standards (Larter, 1991).

Criteria for Selecting a Standard-Setting Method

Ronald Berk's (1986) *A consumer's guide to setting performance standards on criterion-referenced tests* outlines a number of criteria

for selecting a standard-setting method appropriate to the testing situation. The method chosen should:

- yield appropriate classification information given the types of decisions for which the tests are typically used,
- be sensitive to actual examinee performance,
- be sensitive to the instruction that the examinees have actually or potentially received,
- use sound statistical procedures,
- consider measurement error, and
- facilitate valid decision making.

Besides these criteria of a method's technical adequacy, the procedure selected should be practical in that it is easy to implement, to compute, and to interpret to laypeople. Because assessors often have to defend the method, the technique must be credible.

In addition, Berk points to group-interaction and polarization research to suggest that the chosen method should probably be an iterative process. This allows for controlled discussion and permits judges to revise their opinions without making them feel committed to any initial position. The procedure should also allow judges to consider performance data, so they will be less likely to engage in social comparison when setting the standards.

Regardless of a method's statistical elegance or procedural simplicity, all are arbitrary in the sense that there is no scientific procedure that simply involves plugging numbers into a formula. Different methods, as many studies have demonstrated, will produce different standards. But as Popham (1978) has pointed out, castigating standard-setting exercises as unacceptable because they are arbitrary is a non sequitur. The word "arbitrary" can mean either "that which is determinable by a judge or tribunal" or "capricious," that is, "selected at random and without reason." When criticizing standards as arbitrary, critics are clearly employing the second, pejorative definition, when the first definition more accurately reflects serious standard-setting efforts. Standard-setting exercises are judgemental, but to malign all judgemental operations as capricious is absurd (Popham, 1978, p. 168). Even if it is subjective, Livingston and Zieky (1982) contend that "once a standard has been set, the decisions based on it can be made objectively. Instead of a separate set of judgements for each test-taker, you will have the same set of judgements applied to all test-takers. Standards cannot be objectively determined, but they can be objectively applied" (p. 67).

Indeed, standards may relate more to impartial and equitable decision making than to overall educational quality. In the rhetoric of public discourse, standards are often equated with educational improvement. If standards are raised, so the argument goes, then schooling will be better. However, as the Quebec Ministry of Education discovered when it raised the passing scores from 50% to 60% on its Provincial Examinations, one of the principal and unanticipated effects in 1986–1987 was to raise the drop-out rate from its schools (Maheu, 1995, p. 60). Elevating standards does not necessarily translate into better educational outcomes for all students.

Standard-Setting Methods

Although the subjectivity of standard setting is inescapable, some methods are more defensible than others. The traditional approach, and perhaps the least justifiable theoretically, is the random choice of a fixed 50% correct as the standard, without regard to the test's length, difficulty, or importance. Quota-setting involves rank-ordering students on some performance measure such as a test score. Those chosen to participate or receive certification are selected by measuring the number that can be accommodated (usually dependent on the resources or positions available rather than a candidate's actual competence), and counting backwards.

If policy makers have often resorted to one of the foregoing procedures, it has been for want of knowledge about more technically credible alternatives. At least six methods have seen extensive service south of the border, because they are more sensitive to the nature of the actual test, to the criteria used for scoring, to test circumstances, and to the nature of the population being tested. Some reviewers have categorized these methods as either judgemental or empirical; this classification is misleading, however, because all standard-setting methods are judgemental and because many have been modified to include iterations that ask judges to empirically use actual performance data. A more useful distinction is made by methods based on judgements about test items, such as the Nedelsky, Angoff, Ebel, and Jaeger procedures, and those based on judgements about potential examinees, such as the Borderline Group, Contrasting Groups, and Policy Capturing methods.

Nedelsky's method (1954) is restricted to multiple-choice items because the judge's task is to examine each question and identify the wrong answers that a minimally competent test-taker should be able

to eliminate as incorrect. The minimum pass level for each item is equal to the reciprocal of the remaining alternatives. Angoff's method (1971) is similar in that judges make probability estimates for individual items, but it can be used with both multiple-choice and open-response questions. Judges estimate the probability that a minimally competent test taker would answer the question correctly. Modified Angoff procedures have linked the probability estimates to either domain specifications or test objectives, rather than individual test items, to facilitate and accelerate the work of judges.

Ebel's method (1972) differs in that before offering their estimates, judges first rate each question's difficulty (easy, medium, and hard) and its relevance or importance (essential, important, acceptable, and questionable) for the test's purpose. Using these two scales as axes on a matrix, judges classify and group questions into one of the twelve different matrix cells. The total expected score is derived from estimates of examinee performance within each cluster of questions. Variations of this method have substituted levels from taxonomies of higher-order thinking skills along one of the two matrix axes.

One difficulty with these procedures—Berk calls it their "Achilles heel"—is their dependence on judges being able to project themselves into the mental processes of a minimally competent candidate. Unless judges are able to anchor their decision making in evidence and previous experience with youth, they may have the sense that they are "pulling their probability estimates from thin air" (Shepard, 1980, p. 453). In other words, the experts' reasoning processes may diverge substantially from those of the novices they are trying to impersonate, leading to unreliable standards.

Jaeger's method (1978) avoids this difficulty by using an iterative process, employing judges from a variety of backgrounds, providing normative information, and asking empanelled members to consider simply whether every high school graduate should be able to answer the item correctly for a high school diploma. This procedure may be more accessible for lay judges because of the simplicity of the question, but it has been criticized because it restricts probability choices to 1 or 0. A variation of the method (Busch & Jaeger, 1984) was employed for essay responses. Sets of actual papers were presented in random versus ordered and blind versus informed procedures, for judges' estimates. Essays presented in increasing order of quality as scored originally by trained readers, but without prior knowledge of the scores assigned by those readers, allowed judges to produce the most consistent standard.

If the previous methods require judgements about individual test items, both the Borderline Group and Contrasting Group methods, formulated by Livingston and Zieky (1982), require judgements about the students who take the test. The Borderline Group method preselects a subsample of examinees whose skills are marginal; a median test score for this group is chosen as the cutoff score. If the Borderline Group method starts by defining the boundary between competence and incompetence and extrapolating to identify the groups above and below the cutoff score, the Contrasting Groups method works in the opposite direction. Two contrasting groups of clearly qualified and unqualified candidates are preselected before test administration; frequency distributions and percentages are calculated for group test scores and the threshold is determined by interpolating.

Only recently have new procedures been proposed for a complex, multidimensional assessment that encompasses videotaped performances, reflective essays, analytical and critical responses, and divergent planning exercises, all scored with scoring rubrics across a variety of subject areas. To set an integrated standard, Jaeger (1993) has suggested a policy-capturing method in which hypothetical profiles of potential candidates' performances are presented in contrasting sets to judges. Using a four-point scale of competence when rating the profiles, a judgement policy is created for each judge, using a variety of mathematical analyses. Performance standards are derived from the aggregated policies of consistent judges. Considered in light of Berk's four criteria of practicality, the policy-capturing method appears to hold limited promise.

Selecting Judges for Standard-Setting

Many scholars suggest that judges be drawn from different constituencies, so that the standard-setting process can systematically represent "different value positions and areas of expertise" (Shepard, 1980, p. 454; Hambleton & Powell, 1983; Jaeger, 1978) to increase the validity of the recommended standards. Yet few specific guidelines have been proffered either for selecting these judges or for meaningfully incorporating educational stakeholder groups into the standard-setting process. To what degree should judges have expertise in the subject matter being tested, or knowledge of the attributes of the population being tested? Is a panel of classroom teachers preferred over a mixed panel of educators and non-educators? Will a standard produced by a "blue-ribbon" panel be more credible than that produced by an anonymous jury?

Two studies begin to answer these questions. An ill-fated attempt to combine both educators and non-educators on an American standards panel occurred for the 1990 National Assessment of Educational Progress (NAEP) in Mathematics (Bourque & Hambleton, 1993). Competing organizational agendas of stakeholder groups, insufficient training time, a lack of consensus among stakeholders about correct procedures, and a lack of expertise on the part of delegated judges were cited as difficulties. However, when both teachers and university faculty were mixed in a two-stage Angoff procedure (Busch & Jaeger, 1990), and provided both item statistics and actual examinee performance data, judges were "reasoned in their decisions." Although "public school teachers shifted in the direction of the mean standard initially recommended by the college and university judges" (p. 153), possibly because of the prestige associated with the latter group, the difference in standards produced by the two groups was reasonably small.

Indeed, Jaeger (1989a) distinguishes between standards established by policy makers, whose authority to establish test standards derives from position rather than qualification, and those that are based on specialists' judgements concerning the difficulty of test items. For the latter type, it is necessary to recruit judges who have expertise "in the domains being tested and in the roles sought by successful examinees" (p. 3). Experts are distinguishable by their ability to organize and perceive large meaningful patterns in their domains; their skill in analyzing problems rapidly, deeply, and qualitatively; their skills in self-monitoring and judging problem difficulty; and their elaborated semantic memory. Judges should be selected using procedures that permit generalization of their collective recommendations to well-defined populations. The numbers of judges necessary should provide a statistically precise estimation of the standard that would be recommended by an entire population of judges.

Just as there are few guidelines for selecting a heterogeneous sample of judges, there are also disagreements about proper procedures for ensuring these judges produce a reliable standard. Some scholars prescribe an iterative process to minimize the undesirable effect of group domination by a few vociferous members, to have judges privately record colleagues' opinions to balance their individual decision making, and to carefully control discussions to reduce "explicit references of group members to their own positions" (Fitzpatrick, 1984, p. 17). In contrast, others recommend that discussions around individual test items should focus on those judges who diverge most in their ratings from the group's provisional stand-

ard (Arrasmith & Hambleton, 1988; Livingston & Zieky, 1982). Some researchers have suggested that test standards be based on the median distribution of judges' recommendations, rather than the mean, to moderate extreme recommendations. Critics counter that the median is a less stable statistic than the mean (Jaeger, 1989a). Yet other scholars suggest using the trimmed mean—that is, excluding extreme recommendations (Livingston & Zieky, 1982). This practice may be incompatible with, even counterproductive to, the inclusionary purposes of incorporating nonexperts in a process designed to address issues of public accountability.

Two standard-setting approaches that have been used successfully in province-wide Saskatchewan student assessments follow.

STANDARD SETTING IN SASKATCHEWAN

Curriculum Evaluation

Saskatchewan's Curriculum Evaluation program is designed to monitor the effectiveness of new, provincially developed curriculum, to determine levels of curriculum implementation, and to identify strengths and weaknesses that will assist educators in making instructional improvements. In 1993, the first evaluation was conducted with grades 1 to 5 science curricula. Because the program was for purposes of curriculum and instructional improvement rather than public accountability, standards were set by grade-level panels of teachers experienced with the philosophy and content of the curriculum in question. Urban, rural, and northern teachers were selected to provide broad provincial representation.

Initially, the various test items and answer keys were sorted according to the Factors of Scientific Literacy, or skill objectives, identified in the provincial curriculum guide. The teacher-judges then reviewed all materials and discussed at length the expected levels of student achievement according to a five-point scale for each content and skill area assessed. Judges were asked to consider the number of years allowed for curriculum implementation, the amount of direction suggested in the curriculum guide concerning anticipated levels of student mastery, and the assessment items' perceived degree of difficulty.

A department facilitator led standards committee members through a three-stage, iterative process. In the first stage, no actual achieve-

ment results were provided. Instead, based on their personal classroom and curriculum experience, the teachers were asked to consider what minimum scores students at each grade level would have to achieve, on clusters of items representing the various content and skill areas, for performance to be described as strong, very good, satisfactory, marginal, or weak. Descriptions for these performances were elicited as judges discussed potential cutoff scores. Eventually a consensus was reached by each of the grade level committees on the cutoff scores to be provisionally applied in setting standards for each content and skill area.

In the second stage, the cutoff scores were used to calculate the proportion of students that actually achieved each of the five performance levels. These data were then presented to the standards committee, and discussion followed on the relationship between the actual student results and the provisional expectations or standards set in the first stage of the process. To illustrate this process, Tables 1 to 4 show, for four of six Standards for Student Performance (Dimensions of Scientific Literacy) in grade 5 science, the cutoff scores that were derived for each curriculum skill objective and performance category, the actual mean scores students attained for each skill area, and the proportion of classrooms that reached each performance category, based on the cutoff scores. Each assessment item was weighted equally and given a value of 1.0; therefore, the cutoff scores are expressed as decimal fractions.

In the third stage, the empanelled judges were given the opportunity to revise their standards when it was deemed necessary. Again, as the committee considered student results along with the standards or cutoff scores, consensus was reached on final standards. In the Science Curriculum Evaluation project, judges made only two minor adjustments to their initial standards in light of actual student results.

Provincial Learning Assessment

The 1994 Provincial Learning Assessment in Language Arts was designed to provide reliable information to the Department of Education, Training and Employment and the general public about grades 5, 8, and 11 students' skills in reading and writing. Because the program's primary purpose is to address issues of public accountability, and because information is collected for inclusion in a wider education indicators' framework that has been established through

the joint efforts of a wide range of educational stakeholders in the province, standards were established by empanelled grade-level groups of judges designated by the major stakeholder organizations in Saskatchewan's educational system.

The standards committee consisted of delegates from the teaching profession chosen to represent rural, urban, and northern school settings; the Saskatchewan Teachers' Federation; the Saskatchewan Chamber of Commerce; postsecondary institutions including the two universities; the Saskatchewan School Trustees Association; the

Table 1(a)
Standards for Student Performance: Nature of Science

Factor of Scientific Literacy (Skill Objectives)	Student Performance Categories (cutoff scores)[a] Strong (ST)	Very Good (VG)	Satisfactory (S)	Marginal (M)	Weak (W)
A2 Historic	.70	.65	.55	.35	<.35
A3 Holistic	.70	.60	.50	.30	<.30
A4 Replicable	.65	.50	.35	.20	<.20
A5 Empirical	.75	.65	.55	.35	<.35
Nature of Science	.70	.60	.47	.30	<.30

Table 1(b)
Standards for Student Performance: Nature of Science

Factor of Scientific Literacy (Skill Objectives)	Actual Mean W[b]	Actual Mean P[b]	Performance Category (W)	Performance Category (P)	Proportion of Classrooms in Each Category (%) ST W/P	VG W/P	S W/P	M W/P	W W/P
A2 Historic	.44	—	M	—					
A3 Holistic	.57	—	S	—					
A4 Replicable	.53	.29	VG	M					
A5 Empirical	.45	.87	M	ST					
Nature of Science	.53	.58	S	S	6.4 15.4	21.2 36.9	47.4 38.5	21.2 7.7	3.8 1.5

[a] The cut–off score is the minimum score needed to achieve the student performance category.

[b] W refers to student achievement on written assessment tests. P refers to student achievement on performance assessment or practical tasks. Under "Proportion of classrooms in Each Category (%)", the top row of numbers refers to the written assessment tests; the bottom row of numbers refers to the performance assessment or practical tasks.

Table 2(a)
Standards for Student Performance: Key Science Concepts

Factor of Scientific Literacy (Skill Objectives)	Student Performance Categories (cutoff scores)[a] Strong (ST)	Very Good (VG)	Satisfactory (S)	Marginal (M)	Weak (W)
B1 Change	.50	.40	.30	.20	<.20
B2 Interaction	.60	.50	.40	.30	<.30
B4 Organism	.60	.50	.40	.30	<.30
B6 Symmetry	.90	.80	.70	.60	<.60
B7 Force	.70	.50	.30	.20	<.20
B8 Quantification	.75	.65	.50	.40	<.40
B12 Conservation	.70	.65	.60	.45	<.45
B13 Energy–matter	.65	.55	.50	.40	<.40
B14 Cycle	.90	.75	.50	.40	<.40
B16 System	.70	.65	.50	.35	<.35
Key Science Concepts	.70	.60	.47	.30	<.30

Table 2(b)
Standards for Student Performance: Key Science Concepts

Factor of Scientific Literacy (Skill Objectives)	Actual Mean W[b]	P[b]	Student Performance Categories (cutoff scores)[a] Performance Category (W)	(P)	Proportion of Classrooms in Each Category (%) ST W/P	VG W/P	S W/P	M W/P	W W/P
B1 Change	.39	.42	S	VG					
B2 Interaction	.36	—	M	—					
B4 Organism	.50	—	VG	—					
B6 Symmetry	.65	.64	M	M					
B7 Force	.60	.07	VG	W					
B8 Quantification	.52	.87	S	ST					
B12 Conservation	.60	—	S	—					
B13 Energy–matter	.49	—	M	—					
B14 Cycle	.51	—	S	—					
B16 System	.57	—	S	—					
Key Science Concepts	.53	.58	S	S	1.9	10.8	58.4	23.0	5.8
					11.6	44.9	37.7	2.9	2.9

[a] The cut–off score is the minimum score needed to achieve the student performance category.

[b] W refers to student achievement on written assessment tests. P refers to student achievement on performance assessment or practical tasks. Under "Proportion of classrooms in Each Category (%)", the top row of numbers refers to the written assessment tests; the bottom row of numbers refers to the performance assessment or practical tasks.

Table 3(a)
Standards for Student Performance: Processes of Science

Factor of Scientific Literacy (Skill Objectives)	Student Performance Categories (cutoff scores)[a] Strong (ST)	Very Good (VG)	Satisfactory (S)	Marginal (M)	Weak (W)
C1 Classifying	.80	.67	.55	.25	<.25
C2 Communicating	.80	.70	.60	.20	<.20
C3 Observing and describing	.80	.70	.60	.40	<.40
C5 Measuring	.90	.80	.70	.60	<.60
C6 Questioning	.65	.50	.40	.20	<.20
C7 Using numbers	.55	.35	.20	0	N/A
C8 Hypothesizing	.65	.55	.40	.20	<.20
C9 Inferring	.70	.55	.40	.25	<.25
C10 Predicting	.80	.70	.60	.45	<.45
C11 Controlling variables	.65	.50	.30	.15	<.15
C12 Interpreting data	.65	.55	.40	.20	<.20
Processes of Science	.72	.60	.47	.26	<.26

Table 3(b)
Standards for Student Performance: Processes of Science

Factor of Scientific Literacy (Skill Objectives)	Actual Mean W[b]	P[b]	Student Performance Categories (cutoff scores)[a] Performance Category (W)	(P)	Proportion of Classrooms in Each Category (%) ST W/P	VG W/P	S W/P	M W/P	W W/P
C1 Classifying	.48	—	M	—					
C2 Communicating	.49	.91	M	ST					
C3 Observing and describing	.37	.58	W	M					
C5 Measuring	.47	.87	W	VG					
C6 Questioning	.56	—	VG	—					
C7 Using numbers	.62	—	ST	—					
C8 Hypothesizing	.38	.29	M	M					
C9 Inferring	.49	.37	S	M					
C10 Predicting	.54	.88	M	ST					
C11 Controlling variables	.41	—	S	—					
C12 Interpreting data	.56	.35	VG	M					
Processes of Science	.49	.52	S	S	1.3	5.8	60.2	32.1	0.6
					0.0	24.6	60.9	14.5	0.0

[a] The cut–off score is the minimum score needed to achieve the student performance category.

[b] W refers to student achievement on written assessment tests. P refers to student achievement on performance assessment or practical tasks. Under "Proportion of classrooms in Each Category (%)", the top row of numbers refers to the written assessment tests; the bottom row of numbers refers to the performance assessment or practical tasks.

Table 4(a)
Standards for Student Performance: Scientific and Technical Skills

Factor of Scientific Literacy (Skill Objectives)	Student Performance Categories (cutoff scores)[a] Strong (ST)	Very Good (VG)	Satisfactory (S)	Marginal (M)	Weak (W)
E2 Using natural environments	.90	.80	.65	.50	<.50
E3 Using equipment safely	.88	.80	.70	.55	<.55
E7 Manipulative ability	.95	.90	.80	.75	<.75
E10 Measuring temperature	.90	.85	.75	.70	<.70
E11 Measuring mass	.80	.75	.60	.50	<.50
Scientific and Technical Skills	.89	.82	.70	.60	<.60

Table 4(b)
Standards for Student Performance: Scientific and Technical Skills

Factor of Scientific Literacy (Skill Objectives)	Student Performance Categories (cutoff scores)[a] Actual Mean W[b]	P[b]	Performance Category (W)	(P)	Proportion of Classrooms in Each Category (%) ST W/P	VG W/P	S W/P	M W/P	W W/P
E2 Using natural environments	.65	—	S	—					
E3 Using equipment safely	.66	.78	M	S					
E7 Manipulative ability	—	.89	—	S					
E10 Measuring temperature	.77	.93	S	ST					
E11 Measuring mass	.54	.87	M	ST					
Scientific and Technical Skills	.63	.87	M	VG	1.9 14.5	1.9 37.7	20.6 36.2	39.1 7.3	36.5 4.3

[a] The cut–off score is the minimum score needed to achieve the student performance category.

[b] W refers to student achievement on written assessment tests. P refers to student achievement on performance assessment or practical tasks. Under "Proportion of classrooms in Each Category (%)", the top row of numbers refers to the written assessment tests; the bottom row of numbers refers to the performance assessment or practical tasks.

League of Educational Administrators, Directors and Superintendents; and curriculum specialists from the Department of Education, Training and Employment. When selecting delegates, the stakeholder organizations were advised to select representatives from a variety of backgrounds who believed in collaborative decision making and had some experience with youth in the area of literacy, as general qualifications. The 26 members worked in three grade-level subcommittees.

All members were advised that the standards were to be considered, along with actual student achievement results, in the context of the Provincial Education Indicators Program. Standards would highlight strengths and weaknesses in student performance and would suggest improvements in curriculum, public policy, and instructional strategies. The standards would become two-year targets for students, parents, and educational professionals until the next test cycle.

Standards were established for five dimensions of writing and five types of higher-order reading skills, as well as for the reading and writing instruments as a whole at each of the three grades, using a modification of the Angoff method. Because both educational professionals and nonexperts were empanelled, judges based their decisions directly on scoring criteria and examples of actual student work, which had been used for scoring to exemplify the five performance levels. Reading criteria reflected the foundational objectives of impending departmental curricula; writing criteria reflected both nationally and internationally accepted scales of achievement. Copies of test instruments, achievement scales, and scoring bulletins were confidentially mailed to members to aid their preparation in advance of the standard-setting exercise.

Standards were developed in a three-stage, iterative process. In the first stage, facilitators 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. Judges were then asked, “In this skill area, what percentage of the regular stream school population should be able to attain each performance level?” Without consulting others, judges privately wrote down on a tally sheet their preliminary estimates of proportions of students who should attain each of the five levels. Tally sheets were collected, and a mean distribution was calculated and distributed to all judges.

In the second stage, judges were invited to individually 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 attributes of the school population being tested. Once every judge had spoken, a short group discussion was conducted to allow additional viewpoints to be expressed. Members were then given an opportunity to privately 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 judge was again invited to comment in turn on the committee's revised mean distribution and to participate in a short group discussion. Having heard everyone speak, all members were allowed a second opportunity to privately revise their estimates 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 produce a standard for the reading or writing skill area under consideration.

Tables 5 through 8 illustrate the voting patterns of judges participating in the Provincial Learning Assessment exercise for two of the three grades involved and two of ten skill areas under review. The patterns were typical of those for the entire standard-setting session. Although the design of the exercise did not allow individual judges' voting profiles to be mapped, a comparison of standard deviations and mean ratings with actual results permits an analysis of the consistency or consensus achieved.

In judging students' overall writing ability, both the grade 5 and grade 11 subcommittee members' estimates converged across the three rounds of the exercise, as evidenced through the general trend of contraction in standard deviations. Most divergence was evident in adjudicating level 3 and level 4 skill, that is at the mid-to-upper ranges of performance. Presentation of the actual results for the grade 5 subcommittee facilitated even greater consistency in final summative standards; for the grade 11 panel the presentation of performance data allowed for general consistency except for level 3 performance, where the standard deviation increased markedly in round 3. Mean scores at grade 11 also stabilized near the actual

Table 5
Judges' Voting Patterns: Grade 5 Standard–Setting Exercise in Holistic Writing

	Round 1			Round 2			Actual Results[a,b]	Round 3: Final Standard		
	N	Mean	SD	N	Mean	SD	(Standard Error: .034)	N	Mean	SD
% of Students in Level 1	9	10.56	3.50	8	9.25	2.92	5%	7	6.71	2.81
% of Students in Level 2	9	24.56	4.72	8	23.87	2.80	30%	7	26.86	3.53
% of Students in Level 3	9	39.78	9.36	8	41.88	7.75	50%	7	46.29	2.29
% of Students in Level 4	9	18.11	5.84	8	18.62	4.87	13%	7	16.57	3.99
% of Students in Level 5	9	7.00	2.12	8	6.37	1.19	1%	7	3.57	1.99

[a] Level 1 is the lowest and Level 5 is the highest level. [b] 1% of students scored at a level less than 1.

Table 6
Judges' Voting Patterns: Grade 11 Standard–Setting Exercise in Holistic Writing

	Round 1			Round 2			Actual Results[a,b]	Round 3: Final Standard		
	N	Mean	SD	N	Mean	SD	(Standard Error: .038)	N	Mean	SD
% of Students in Level 1	8	9.88	2.70	9	8.33	2.92	3%	9	3.89	3.06
% of Students in Level 2	9	15.33	5.57	9	16.89	3.48	20%	9	17.56	4.82
% of Students in Level 3	9	35.36	5.77	9	37.22	2.82	54%	9	47.67	8.90
% of Students in Level 4	9	25.67	7.26	9	26.11	6.81	20%	9	23.22	5.56
% of Students in Level 5	9	13.78	3.31	9	11.33	3.46	1%	9	7.22	2.82

[a] Level 1 is the lowest and Level 5 is the highest level. [b] 2% of students scored at a level less than 1.

results except for performance level 3, which witnessed an 11-point adjustment in the mean standard between round 1 and round 3.

When making decisions about student reading comprehension competence, judges' voting patterns for grade 5 did converge through discussions between rounds 1 and 2. But the presentation of actual student achievement results in round 3 had the opposite effect for the grade 5 subcommittee members; standard deviations increased markedly for estimates of what students should be able to achieve for the top four performance levels as did the mean standards proffered by subcommittee members. On the other hand, the overall voting patterns of grade 11 judges stabilized across the three rounds, and standard deviations generally narrowed except for performance levels 3 and 4.

Taken together, voting patterns suggest that the procedure employed for the Learning Assessment did facilitate a general convergence of stakeholder judgements for provincial literacy standards. However, the presentation of actual results sometimes resulted in diverging judgements, particularly for levels 3 and 4 performance. The heterogeneous panel of judges could generally reconcile expectations for lower- and upper-range competence, but had difficulty adopting a shared view of what should constitute mid-range competence. Stable voting patterns did not emerge in several instances after actual performance data were presented.

Nevertheless, overall patterns of convergence indicate that empanelled stakeholder representatives were collaborative and consensual in their decision making. The adjustments in scores to reflect actual normative data suggest that judges based their decisions on information provided by the standard-setting procedure. That the overall standards approximated a normal distribution curve suggests that judges were sensitive to the typical range of student ability. That grade 5 judges permitted actual reading comprehension results at the upper ranges to exceed their desired performance standards demonstrates a reasoned approach to defining what students should be able to do.

DISCUSSION

Although the two procedures employed in Saskatchewan were both variations of Angoff's method, a number of differences are apparent. The Provincial Learning Assessment exercise proceeded on the

Table 7
Judges' Voting Patterns: Grade 5 Standard–Setting Exercise in Reading Comprehension

	Round 1			Round 2			Actual Results[a]	Round 3: Final Standard		
	N	Mean	SD	N	Mean	SD	(Standard Error: .034)	N	Mean	SD
% of Students in Level 1	8	10.88	5.87	9	10.78	3.15	4%	6	5.67	2.16
% of Students in Level 2	8	18.87	9.20	9	20.01	5.01	15%	6	20.83	9.70
% of Students in Level 3	8	36.25	15.30	9	38.67	9.42	20%	6	34.33	10.42
% of Students in Level 4	8	22.00	9.27	9	19.89	5.33	30%	6	24.17	8.01
% of Students in Level 5	8	12.00	9.35	9	10.67	4.27	31%	6	15.00	11.83

[a] Level 1 is the lowest and Level 5 is the highest level.

Table 8
Judges' Voting Patterns: Grade 11 Standard–Setting Exercise in Reading Comprehension

	Round 1			Round 2			Actual Results[a]	Round 3: Final Standard		
	N	Mean	SD	N	Mean	SD	(Standard Error: .038)	N	Mean	SD
% of Students in Level 1	8	7.00	3.51	9	8.89	3.18	3%	8	4.62	0.74
% of Students in Level 2	9	15.67	7.68	9	15.89	5.04	14%	9	15.22	5.59
% of Students in Level 3	9	27.00	9.34	9	31.78	13.98	40%	9	39.67	6.24
% of Students in Level 4	9	31.00	7.84	9	27.56	8.89	38%	9	30.56	8.83
% of Students in Level 5	9	19.33	5.83	9	16.00	6.75	5%	9	10.89	3.76

[a] Level 1 is the lowest and Level 5 is the highest level.

basis of performance levels and criteria that had been established prior to scoring, and the Curriculum Evaluation project used the standard-setting exercise as a vehicle for defining the performance descriptions. Both procedures worked well. Although the former anchored standards more directly in scoring judgements, the latter was deemed more satisfactory to judges: giving the judges the task of not only establishing expectations but also defining the performance levels enabled them to develop more ownership of the finished product.

More importantly, the two procedures employed two different notions of consensus. Because of anticipated divergent expectations between educator and non-educator judges for the Provincial Learning Assessment, the procedure aimed only at deriving a mathematical merging of judges' expectations rather than an actual *consensus ad idem* or meeting of all the minds involved. On the other hand, the Science Curriculum Evaluation exercise was predicated on actual group consensus of the empanelled judges; a hung jury was not allowed. In general, judges in the Learning Assessment believed that a mathematical consensus without simultaneous social consensus tended to undermine the validity of the standard.

The pivot around which all standard-setting procedures revolve is the question to which judges are to respond. For the Saskatchewan Learning Assessment exercise, facilitators asked about the percentages of students who *should* be able to attain the desired levels of performance; the Curriculum Evaluation exercise hinged on the percentages of students who *would* be able to attain each level. The first asked judges to conceptualize, in the ideal, what the potential is for student performance, but the second asked for a formulation, in light of the reality of classroom experience, of actual student competence. In a sense, the former defined expectations as targets, whereas the latter defined expectations as thresholds of acceptability. The effect of varying questions on the resultant standard warrants further investigation. Having an empanelled jury answer both questions, and noting commentary on individual ballots to explain the discrepancy in their estimates, may facilitate interpretation.

In both Saskatchewan projects, judges were selected from constituencies that did not have a detailed knowledge of the assessment design, test construction assumptions, or scoring procedures prior to the exercise. Contamination of judgement through experience with preceding assessment phases was deemed possible, leading to unreliable standards. However, for both exercises, a considerable

amount of time was spent briefing judges about the purposes, premises, and preceding phases of the projects. Securing the judges' endorsement of or even acquiescence to these premises and processes of standard-setting exercises may prove disruptive and time-consuming, and may subsequently colour the essential task of judging student performance. Future projects will explore this issue of cross-contamination or cross-fertilization through continuity of membership between the different tasks of scoring and judging student performance.

For both standard-setting exercises, judges were deliberately selected from remote and northern settings, as well as from Aboriginal communities. These perspectives were brought to the adjudication of student performance so that the resultant standard was culturally and geographically sensitive. For the Provincial Learning Assessment exercise, however, concerns were expressed that the standard-setting procedure did not reflect the traditional practices of Aboriginal peoples. In Saskatchewan, the justice system has made adaptations to its court procedures, through sentencing circles, for high-stakes judgements about young adult offenders. Future projects may adapt the standard-setting procedure to customary Aboriginal practices for making "low-stakes" judgements about educational performance.

Evaluating the Effectiveness of Standards-setting Methods

An educational standard-setting procedure can be conceived and evaluated from a variety of perspectives: as a "due-process" mechanism for having judges make reasoned and informed decisions about student competence; as a procedure for measuring the match between what educators intend to do and what they actually have done; or as an instrument of policy for drawing together competing stakeholder interests and values to define systemic performance. The confidence one might have in a performance standard may be determined through examining both the internal consistency of judges' voting patterns—as provided above—or through comparisons with external criteria (Kane, 1994).

Considered in light of Berk's (1986) selection criteria:

- Both the Science Curriculum Evaluation and the Language Arts Learning Assessment procedures yielded appropriate classification information given the differing purposes of the assessments.

- Both methods used in Saskatchewan were sensitive to actual examinee performance because performance data for the identified skills were shared with the panelists to inform their setting of final standards. In addition, item difficulty statistics were provided to panelists for the reading component of the Learning Assessment exercise.

- The Curriculum Evaluation exercise was more sensitive to current or potential instructional practices, because panelists were practising teachers who represented all regions of the province. On the other hand, some standard-setters in the Learning Assessment exercise felt that the method did not adequately attend to the provincial circumstance where language arts curricula are in a state of transition.

- Both exercises employed statistically sound procedures, through simple means and percentages, which accurately represented both the actual profiles of student achievement and the collective expectations of standard-setters.

- Judges' consideration of standard errors of measurement in both Saskatchewan exercises was unnecessary given the low-stakes purposes of the assessments, although panelists in the latter were provided with data about possible sampling error and indices of scoring consistency.

- Both methods facilitated valid decision making. By providing a rating of classroom achievement in light of curricular expectations, decision makers were provided with useful information about elementary science curriculum implementation and impact; the Learning Assessment exercise yielded nuanced information about strengths and weaknesses in student literacy skills, in light of stakeholder opinion, for public accountability and instructional planning purposes.

To summarize, the standard-setting method used in the grade 5 science curriculum evaluation was practical and easy to implement and compute. It produced the required information, and the panelists were very comfortable and satisfied with the process. In the Language Arts Learning Assessment exercise, all judges believed that input from the various parties and organizations was valuable and useful. However, some panelists believed that the procedure would be difficult to explain to those not present.

CONCLUSION

Quite clearly, more work remains in the Canadian setting for this crucial evaluative activity. More research is required into the various types and sources of evidence upon which judges make their decisions, and the relative effectiveness of various standard-setting methods in supplying judges with that information in their decision making. Standard-setting methods must be sensitive to the various cultural perspectives that judges bring to the table, and must be clearly explained to judges before they arrive at the standard-setting session. Agreement as to the assessment design, criteria employed in scoring, and specific standard-setting procedures must be secured before embarking on the task at hand. Thorough orientation and training of judges in the procedures is a prerequisite.

Because education is a provincial responsibility in Canada, provincial ministries of education must further explore methods for establishing student achievement standards for provincial and national assessments. This must be done if questions related to public accountability in education are to be adequately addressed.

REFERENCES

Alberta Education, Student Evaluation Branch. (1991). *Achievement testing program, provincial report: June 1991 administration: Grade 3 science, grade 6 mathematics and grade 9 social studies*. Edmonton, AB: Author.

Alberta Education, Student Evaluation Branch. (1993). *Achievement testing program, provincial report: June 1993 administration*. Edmonton, AB: Author.

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

Arrasmith, D.G., & Hambleton, R.K. (1988). *Steps for setting standards with the Angoff method* (Laboratory of Psychometric and Evaluative Research Rep. No. 168). Amherst: University of Massachusetts.

Bateson, D., Anderson, J., Brigden, S., Day, E. Deiter, B., Eberle, C., Gurney, B., & McConnel, V. (1991). *British Columbia assessment of science 1991. Technical report I: Classical component*. Victoria: British Columbia Ministry of 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), 41–48.

Busch, J.C., & Jaeger, R.M. (1984). *An evaluation of methods for setting standards on the essay portion of the National Teacher Examinations*. Paper presented at the joint annual meeting of the American Educational Research Association and the National Council of Measurement in Education, New Orleans.

Busch, J.C., & Jaeger, R.M. (1990). Influence of type of judge, normative information and discussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement, 27*(2), 145–163.

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

Fitzpatrick, A.R. (1984). *Social influences in standard-setting: The effect of group interaction on individuals' judgements*. Paper presented at the annual meeting of the American Psychological Association, New Orleans.

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

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*(2), 237–263.

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, Chapel Hill, NC.

Jaeger, R.M. (1989a). Selection of judges for standard setting. *Educational Measurement: Issues & Practice, 10*(2), 3–14.

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

Jaeger, R.M. (1993). *Integrating multi-dimensional performances and setting performance standards*. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Jeroski, S. (1989). *The 1988 British Columbia assessment of reading and writing expression: Technical report*. Victoria: British Columbia Ministry of Education.

Joint Advisory Committee. (1993). *Principles for fair student assessment practices for education in Canada*. Edmonton, AB: Centre for Research in Applied Measurement and Evaluation, University of Alberta.

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

Larter, S. (1991). *Benchmarks: The development of a new approach to student evaluation*. Toronto: Toronto Board of Education.

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

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*, 3–19.

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

Shepard, L. (1980). Standard setting issues and methods. *Applied Psychological Measurement, 4*, 447–467.