Darryl M. Hunter
Saskatchewan Department of Education

and

Bikkar S. Randhawa
University of Saskatchewan

The Large-Scale, Authentic Assessment of Listening and Speaking as Interactive Communication: Issues in Reliability

In this article we discuss reliability issues in the large-scale assessment of speech communication through authentic or performance techniques used recently in Saskatchewan. New performance-based approaches enable educators to evaluate the integrated, interpersonal communication skills of large populations of students, thereby modeling best professional practice. We conclude, however, that decentralized teacher rating approaches do not yet demonstrate sufficient reliability to enable public officials to use the results for high-stakes testing purposes.

Dans cet article, nous discutons de questions de fiabliité dans l'évaluation à grande échelle de la communication verbale par le biais de techniques authentiques (ou basées sur la performance) employées en Saskatchewan. De nouvelles approches reposant sur la performance permettent aux enseignants d'évaluer les compétences communicatives intégrées et interpersonnelles de grandes populations d'élèves, réflétant ainsi les meilleures stratégies pédagogiques. Nous concluons toutefois que les approches impliquant une évaluation décentralisée ne s'avèrent pas encore suffisamment fiables pour que les fonctionnaires se servent des résultats dans des contextes où l'enjeu est considérable.

The authentic assessment movement is causing ferment both in measurement circles (Terwilliger, 1997) and in the wider educational community (Stiggins, 1995). Many educators are beginning to regard traditional multiple-choice testing as inadequate for describing student growth and achievement in complex domains such as written expression and mathematical problem-solving. Instead of using norm-referenced tests, assessors are opening new frontiers with criterion measurement through portfolios, performance tasks, and group problem-solving (Stiggins, 1991). As views of learning and evaluation change from behavioral to cognitive and constructivist foundations, an accompanying shift is occurring from paper-and-pencil measures to open-response formats, from single-attribute to multidimensional assessment, and from a near exclusive emphasis on testing individuals to group evaluations. Performance assessment is thus reorienting teachers in their approach to student evaluation

Darryl M. Hunter is Director of Assessment and Evaluation.
Bikkar S. Randhawa is Professor Emeritus of Educational Psychology and Special Education, College of Education.

while also opening new areas for skill measurement through large-scale assessment (Wiggins, 1993a).

Many educators use the terms *authentic assessment* and *performance assessment* as if they were synonymous. Performance measures or tasks (Dunbar, 1991; Wiggins, 1993a) have several properties: that student production should occur through open-ended application of knowledge; that tests must encourage construction of complex performances rather than selection from prepared alternatives; that higher-level cognitive or problem-solving skills are demonstrated; that process behavior is gauged as well as the product; and that individuals or groups should actively use familiar equipment to demonstrate proficiency rather than offer passive responses in pencil-and-paper situations. Performance assessments, moreover, should model good instruction and exemplify those tasks relevant to the curriculum (Wiggins, 1989). Authentic assessment carries the additional qualification that the tasks or outcomes should have real-world connections to the workplace or adult life. Hence contextual relevance or "ecological validity" is an important property of authentic assessment (McLean, 1990; Wiggins, 1993b). As such, all authentic assessment measures are also performance tasks, but not all performance tasks are necessarily authentic. Regardless of which label one uses, both authentic and performance measures present the actual task to be performed and do not just represent it in text form (Randhawa & Hunter, 2001).

A central and unacknowledged presumption behind such measures is that to be valid and useful a test should reproduce the behaviors it purports to measure. If the purpose of testing is to provide information back to participants about their relative standing, then the crucial reliability and validity check is the degree to which the variability of scores accurately matches the variability that actually exists in performance with the construct being measured. Although it is likely that the test will measure aspects of the construct under investigation such as listening and speaking competence, it may not be necessary to do so authentically, but only efficiently and effectively to meet its purposes. On the other hand, if the primary purpose of assessment is to promote curricular adoption, then authenticity may need to be incorporated in the assessment so that teachers "teaching to the test" will thereby implement those instructional practices deemed desirable in prescribed curricula (Bateson, 1994). So too does involving teachers in marking exercises assist with cultivating professional expertise when the purpose of assessment is to build systemic capacity.

In fact, large-scale assessments generally have multiple and sometimes contradictory purposes—to redirect curriculum content and pedagogy, to model best practice for students and teachers, and to report publicly on student or school performance for accountability concerns—which may necessitate trade-offs in their authenticity and in the degree of desirable reliability sought. For some measurement specialists (Frederiksen & Collins, 1989; Moss, 1994) the purposes of assessment and especially the overarching goal of system improvement will determine an acceptable index of reliability. Moreover, Moss argues that conventional notions of reliability need to be reconceived to address not only the traditional concerns with interrater and intrarater reliability, but also the interpretive congruence between third-party testing and classroom evaluation. There may be times when one sacrifices a degree of authenticity, such as with high-stakes testing, in exchange for a degree of reliability (Brennan & Johnson, 1995). Conversely, there may be times, such as with program evaluation, when one barters some internal consistency for a more diffuse pedagogical impact. But these trade-offs need to be deliberate and ought to be considered as short-term exigencies. Eventually, the assessor ought to be able to devise assessment procedures that are at once authentic and consistent (Schippman, Prien, & Katz, 1990; Swanson, Norman, & Linn, 1995).

For others (Nichols & Smith, 1998) the underlying theory of the construct being measured and of learning embedded in the test-takers' elicited performance will determine appropriate indices of reliability. Current methods for quantifying reliability, they contend, privilege theories of learning other than those that underpin cognitively complex, performance assessments. Classical notions of reliability such as test-retest, alternate forms, and internal consistency or its extension in generalizability theory (Brennan, 2000) are often inadequate for ascertaining the consistency of the new authentic assessment approaches that proceed from different assumptions about learning and its measurement.

Accordingly, the purposes of this article are threefold. First, we explore issues in reliability with authentic measurement as contrasted with issues of reliability using more conventional measurement approaches. Second, we describe a large-scale oral assessment that was implemented in one Canadian public school setting and analyze its design in terms of its measurement consistency. We report on the development of new measures of communicative competence in interactive situations and depict the extension of criterion performance rating scales into interactive communication. Third, we suggest new approaches to evaluating the consistency and hence the technical adequacy of authentic or performance measures. In so doing our goal is to enhance the reliability of naturalistic observation as an assessment procedure for its applicability in either the first- or second-language learning situation.

Reliability Through Reintegration of the Attributes Being Assessed

The authentic assessment movement has countered the traditional measurement strategy of separating and rendering distinct the specific skills that are the subject of investigation. Through segmentation and control, the classical measurement specialist strives to ensure stability in measurement of the construct. This has led to fragmentation of skills in testing situations such that the measured properties bear little correspondence to how interpersonal communication typically occurs (Morreale & Backlund, 1996). Against this disintegrative effect, the authentic assessor creates complex, integrated tasks that elicit students' whole performances in open-ended formats that match curricular demands.

To ensure curricular fidelity and content validity in the 1998 Saskatchewan Learning Assessment, a 14-member program team composed of classroom teachers, school administrators, language arts and measurement specialists from the Universities of Regina and Saskatchewan, and curriculum writers from the Saskatchewan Department of Education oversaw the development of all materials. The assessment was designed to fulfill the program purposes of creating provincial-level indicators for public accountability, enhancing professional skills in student evaluation, and assembling a time-series profile of student achievement (Saskatchewan Education, 1999a). Information was to be aggregated and reported on a provincial basis only in terms of criterion-referenced standards; no information was to be reported for individual students, classrooms, schools, or school divisions. Fifteen Saskatchewan teachers developed tasks and rubrics during the fall of 1997 using a table of specifications established by the program team that directly linked the assessment criteria and procedures to the foundational objectives of new Saskatchewan language arts curricula. All materials were subsequently reviewed and revised by the program team, three focus groups of teachers, language arts consultants, and a validation committee of subject area specialists and curriculum specialists. During the first two months of 1998, the materials were field-tested in 13 Saskatchewan schools including rural, urban, and northern locations by the teacher coordinating the project. At these sites, assessment materials were reviewed for clarity and procedures were refined for applicability. Materials and procedures were further refined and streamlined on the basis of this field-testing.

When developing and refining these performance tasks and procedures, several key premises were adopted. First, listening and speaking were conceived as an integrated, interactive act of communication, not as a set of discrete skills (Mead, 1982). Oral communication would involve both verbal and nonverbal exchanges with an audience. The communicative content encompassed not only thoughts or information, but also attitudes. Communication was to occur in a social context: listening and speaking are means by which individuals make connections with each other. To speak fluently and confidently in a variety of situations and to appreciate the needs and positions of others were deemed necessary attributes of effective oral communication.

Second, the communicative task would be presented as a classroom instructional activity, not represented in pencil-and-paper and audiotape formats. Students would be prompted in sequence with two tasks that asked small groups of students to interact in a manner that reflected a typical, cooperative learning response to a real-world problem. Rather than isolating the student in front of his or her audience or in a listening booth with earphones and cued cassette tapes, the provincial assessment asked students to watch a televised video clip prompt and subsequently to discuss in a small group the moral and social issues raised by the clip. In grade 5 students were asked to discuss the qualities of people they admired, in grade 8 issues of allegiance to friends, and in grade 11 the social ethics questions involved in AIDS and teen pregnancy. Issues were framed as a single multidimensional task or problem for resolution, with a subsequent class presentation of the group's solution and its rationale.

The third presupposition was that a variety of communicative competences could be descriptively synthesized and incorporated into holistic rating scales that conveyed a variety of primary traits in six communicative dimensions (Rubin, 1981). These criterion rubrics, with accompanying videotaped illustrative performances, addressed three aspects of group discussion for assessment of listening behavior (Rubin & Roberts, 1987). The first dimension was *participation*, which assumed that student involvement is shaped by a number of factors, including communicative intent or purpose, assignment focus, social structure, and language. Group members' willingness to voice ideas and opinions and to share experiences was deemed to be a part of participation. This participation would be measured in terms of both the quality and quantity of participation. Second, *active listening skills* were gauged: the verbal and nonverbal communicative behavior of others were to be acknowledged and built on. Raters had to attend to the posture and facial expression of examinees, in keeping with their purpose, and with the cultural and linguistic background and needs of others. And third, those traits or attributes that demonstrated *respect for conversational peers* were described in rubric form. Proficient communicators convey their awareness of and concern for other members in a group setting by showing tact in the expression of their ideas, polite interjections, and respect for ideas and opinions different from their own.

Similarly, holistic rating scales were developed for assessment of speaking skills in a group presentation. The quality of communicative behaviors was categorized in three dimensions (Rubin, 1985). First, the substantive aspects of the message were covered under the label of *content*: both the quality of ideas and information presented, as well as of the organization of the communicative substance, were described in graduated performance levels. Second, competence in language use was ascertained by looking at the appropriateness of language choices for the specific speaking situation, as were the clarity of the speaker's language for conveying meaning and the speaker's ability to use language to create a unique personal style. The third rubric captured *presentation or delivery style*, including voice and bodily behaviors that accompany the speaker's words, including inflection, articulation, rhythm, facial expression, eye contact, and gestures.

Each of these six rubrics was graduated, from Level 1 (low) through 4 (high) and considered to reflect stages on a continuum (Saskatchewan Education, 1999a). The scales incorporated primary traits that differentiated through the precise choice of qualitative adjectives and adverbs rather than through the quantitative presence or absence of behaviors (Quellmalz, 1991). The development of listening and speaking skills in a group setting was assumed to begin well before Level 1 and to extend beyond Level 4. These performance scales captured holistically the primary traits of speech communication: not every attribute in the rubric needed to be demonstrated to identify a student at a particular criterion level.

Reliability Through Preparation of Teacher Scorers

Whereas the classical measurement specialist will remove potential vagaries by sharply restricting the judgment of teacher practitioners, the authentic assessor repositions the teacher along with his or her criterion tools as the central agents in large-scale testing. In the 1998 Saskatchewan study, teachers were asked to rate their students as small groups in the regular language arts classroom. A third-party agency would not intrude directly in the school situation or score student work in a decontextualized marking room. In contrast, scores would be assigned by a participant in the communicative process (Webb, 1995). Just as the multidimensional communicative task synthesized a variety of appreciative, cognitive, and social skills in a communicative situation, teachers were asked to make integrated appraisals when assigning a rating. Rather than basing judgments in associationist or elemental theories of perception—which see the parts creating a whole—Gestalt psychologists see the whole as determining the character of its parts (Hunter, Jones, & Randhawa, 1996). Hence teachers in the Saskatchewan assessment were asked to use the textual Gestalts found in the six rubric rating scales as *grounds*, and exemplar performances in video form as *figures*, as criterion referents for rendering their judgments. The purpose of preparatory training was to reorient teachers to see and appraise the communicative whole and its figurative expressions consistently in the classroom.

In early April 1998, the 87 classroom teachers sampled were invited to attend one of four two-day orientation sessions held between April 27 and May 8, 1998. The orientation familiarized teachers with the Listening and Speaking Assessment premises and procedures and provided intensive training for scoring students' performance. During the orientation teachers considered common rater errors, became familiar with the scoring rubrics, and practiced rating with videotaped illustrations of group work using the six scoring rubrics.

Reliability training included careful review of the criteria in the six dimensions and large-group discussion to develop a shared interpretive outlook among teachers in applying these criteria. Limiting the number of raters trained in any one session allowed all raters the opportunity to engage actively in the discussion following the practice ratings. Videotaped student performances were presented—with a wide range of quality levels for both group discussions and group presentations—including those that easily elicited consensus in teacher ratings and those that provoked varied responses so as to provide teachers with a more thorough training. Adjudication was provided when necessary by a scoring leader during practice rounds to enable scorers to anchor consistently into the rating levels described in the rubrics.

Reliability Through Reconfiguration of Assessment Design

A sometimes overlooked aspect of reliability in large-scale assessment derives from the overall project design. How stable is the administrative and procedural architecture? If measures of oral communicative competence are to be consistent, they must be administratively feasible and amenable to uniform application across diverse situations. Considerations such as the amount of student time allocated; consistent allocation of personnel for administration and scoring; and the requirement of specialized training for administration, scoring, and interpretation all affect the reliability of large-scale assessment, particularly for naturalistic evaluation procedures. Therefore, tests of competence, particularly with speaking, need incorporate cost-effective means for capturing and judging a performance. An effective listening or speaking assessment must be a reliable measure not of administrative commitment, but rather of the oral competence of students.

Whereas the traditional test often involves the imposition of an invariable and preprogrammed procedure onto the "natural" classroom or school environment, the authentic assessor strives to collect information that fits flexibly in the school ecology to minimize its perceived intrusiveness on teaching and learning. Both the authentic assessor and the classical measurement specialist share an identical goal of creating an internally consistent and similarly structured situation wherein information is collected. Whereas the traditional approach would be to exert centralized environmental controls, the authentic assessor aims to decentralize and embed a complex task consistently in a larger social context that may vary from one school situation to another.

The sampling design for the 1998 assessment was quite conventional. Participating schools were randomly selected from the list of provincial schools teaching the Saskatchewan curriculum with grades 5, 8, and 11 students. Selection criteria produced a random 8% student sample totaling 1,967 pupils (608 in grade 5; 686 in grade 8; 673 in grade 11) that proportionally represented rural, urban, and northern schools. Schools of varying population sizes were represented. Criteria excluded schools that had field-tested the assessment instruments and small schools involved in parallel national or provincial reading and writing assessments. Schools in francophone school divisions did not participate. Schools with fewer than 30 students in the chosen grade had all eligible students participate. Schools with more than 30 students selected a class of at least 25 students.

Only those students enrolled in regular English language arts programs were involved. Students in modified programs where foundational objectives of the course had been altered did not participate. In other words, students in special resource room programs or International Baccalaureate (IB) programs or in locally developed courses were not assessed. Selected schools and their school divisions were contacted in February 1998. Information bulletins outlined the nature and scope of the assessment for teachers, parents, and students. Class lists were submitted to the Department of Education for random assignment of students into small groups of four or five members for the actual assessment.

Allowances were made for adaptations such as large-print forms, scribes, or other special needs. Administrative guidelines also aimed to ensure consistent administration of the assessment in terms of time allocations, breaks, independent student work, absenteeism, the type of assistance provided, and rigorous scoring of student performance. The guidelines also sought to ensure appropriate handling of student work to guarantee student, teacher, school, and school division anonymity. The actual assessment took place between May 11 and May 22, 1998. A total time allotment of 3.5-4 hours was provided for conducting the assessment in each classroom.

Structurally, the assessment of listening and speaking proceeded in four stages following a blueprint somewhat different than typical large-scale standardized testing. First, after viewing a video prompt, students' individual listening abilities were assessed using a 15-question multiple-choice test in written form with questions that were read aloud to students. This closed-response exercise served as a measure of concurrent validity and is not reported in this study. In the second stage students were given a group discussion task in both oral and written form to complete in their preassigned small groups. They were asked to participate in a small-group discussion to prepare a 2-4-minute group presentation describing their collaborative response to the assignment. The student test booklet and oral instructions guided this discussion: these included preparatory questions, a planning chart, and a checklist. During the small-group discussion the pretrained teacher rated each group for participation, active listening, and respect for conversational peers. In the third phase each student group delivered its 2-4-minute presentation to their classmates with their solution to the given assignment. The same pretrained teacher rated each small group as a unit for content, language use, and presentation or delivery style. In the fourth stage students completed: a self-evaluation form to rate their individual performance in the group performance; a peer evaluation form to assess how their group had performed in both the discussion and presentation phases; and a student questionnaire about their speaking and listening behaviors and attitudes.

Thus in adopting authentic rather than classical measurement assumptions, the 1998 Saskatchewan Listening and Speaking Assessment entailed a new conceptualization of the construct under consideration, a reorientation in the teacher's role vis-à-vis third party testing, and a different administrative design for the large-scale initiative. Our central question is, then, how reliable was the measurement of students' listening and speaking skills as interactive communication?

Analysis

Appropriate data for the three reliability facets described in this article were used to compute as necessary simple correlation coefficients, means, standard deviations, and percentage agreements. Whereas a correlation coefficient between two measured variables indicates the degree of association or consistency, means and standard deviations provide descriptive properties of the variables. From these descriptive measures one can assess the relative similarity or differences of averages (means) and whether the scores had similar or different spreads (variability). As the situation warrants, the size of percentage agreement provides an index of consistency of different raters or a rater's assigned value with a target or standard value.

Findings

Reliability in Integrating the Attributes Assessed

To investigate the consistency with which teachers were able to integrate holistically the various aspects of listening competence in the group discussion situation, intercorrelations were calculated for those scores assigned at the grade 8 level and are shown in Table 1. Mean scores and standard deviations on the 4-point scale are also provided in this table as complementary data. The

Table 1
Intercorrelations, Means, and Standard Deviations of Group Discussion (Listening) Dimensions as Rated by Teachers, 1998 Saskatchewan Listening and Speaking Assessment (N^a=193)

Dimension	*1*	*2*	*3*
1. Participation	1.00	.72	.66
2. Listening	.72	1.00	.69
3. Respect	.66	.69	1.00
Mean	2.58	2.57	2.77
SD	.83	.89	.85

[a]Represents number of grade 8 student groups involved in assessment.

three listening dimensions must be considered as separate from the succeeding speaking dimensions because the task situations and prompts were different, and because ratings were assigned in a different temporal context.

The mean scores on the three dimensions are 2.6, 2.6, and 2.8 respectively, which are almost equivalent—that is, there are no statistically significant differences between the pairs of means. The standard deviations of the three dimensions are .83, .89, and .85 respectively, which again are statistically similar. The correlation coefficients range from .66 to .72, indicating moderately high redundancy between the pairs of dimensions in listening assessed in this project. Also, the correlation coefficients of the three pairs of the three listening dimensions are statistically similar. This may mean either that students exhibited generally the same range and quality of listening skills regardless of dimension, or that teachers were generally applying the same underlying units of measurement to the three dimensions of group discussion. Alternatively, and perhaps more plausibly, the statistical similarities may indicate that students' listening competences are related to a more general listening ability.

Table 2 provides a measure of the internal consistency of measurement by teachers for three dimensions of speaking competence: content, language usage, and presentation or delivery style. Also included in this table are the mean scores and standard deviations for the ratings on the three speaking competence dimensions assigned by teacher-scorers.

Here the picture is similar to that for listening. The observed mean scores are almost the same, with values of 2.3 when rounded. However, the standard deviations for these dimensions are .93, .77, and .92. It appears that usage ratings had smallest variability compared with the substantive content and stylistic elements of speech communication. This lower variability of ratings for linguistic usage suggests that classroom teachers viewed student performances through a narrower aperture.

The correlation coefficients for the three pairs of speaking dimensions are statistically similar at .58, .62, and .65: they reveal either that teachers had difficulty discriminating between the elements or that they were successful in integrating the various elements of speaking when rendering their judgments of students' group performances. The highest amount of redundancy (42%)

Table 2
Intercorrelations, Means, and Standard Deviations of Group Presentation (Speaking) Dimensions as Rated by Teachers, 1998 Saskatchewan Listening and Speaking Assessment (N^a=189)

Dimension	*1*	*2*	*3*
1. Content	1.00	.58	.62
2. Language	.58	1.00	.65
3. Presentation	.62	.65	1.00
Mean	2.28	2.30	2.32
SD	.93	.77	.92

[a]Represents the number of grade 8 student groups involved in assessment.

was between the quality of language usage and ratings of delivery style (presentation), whereas the lowest amount of redundancy (34%) was between the substantive material (content) and ratings for level of linguistic sophistication (language usage).

As for the preceding listening dimensions, teachers used a generally stable framework of appraisal when judging competence in speaking. This may mean that teachers were consistent in their measurement approach despite the remarkably different school contexts in which the assessment was conducted across the province and the different ways that groups of students may have responded to the prompt. Alternatively, it may mean that teachers viewed group performances as being relatively homogeneous in quality regardless of the scoring rubric deployed.

Reliability of Teacher Raters

Whereas Tables 1 and 2 illustrate the degree of reliability in examining the attributes under investigation, Table 3 shows the levels of agreement of teacher agents in collecting information. Immediately following their training sessions, all 87 raters across three grade levels were given two separate videotaped performances, in both group discussion and group presentation, at their appropriate grade levels. Inter- or intrarater reliability coefficients were not computed because of their inapplicability in this assessment design. Rather, stability in measurement with the centrally predetermined scoring benchmarks is examined through rater consistency immediately following their training and subsequently through agreement with a rating supplied by an independent reviewer who entered the classroom as an external reliability check.

As Table 3 reveals, between two thirds and nearly three quarters of teachers assigned an identical rating for the group discussion dimensions during the first quality control check and similar percentages during the second check. Reliability declined between the two checks for the participation dimension from 73.6% to 60.9%, but increased for the active listening and respect dimensions by 7-10%. This indicates that between 35% and 40% of the teachers were

Table 3
Rater Agreement[a] with Target Rating Following Centralized Training, All Teachers, 1998 Saskatchewan Listening and Speaking Assessment (*N*=87)

	Check 1	*Check 2*
Group Discussion Dimensions		
Participation	73.6	60.9
Listening	60.9	67.8
Respect for peers	60.9	71.3
Group Presentation Dimensions		
Content	50.5	62.1
Language	57.5	74.7
Presentation style	71.3	63.2

[a] Reliability statistic represents the percentage of teachers who agreed with a target score for a videotaped performance pre-rated by the assessment coordinator.

not calibrating with the centrally determined standards at the end of the training session for the listening task and its judgmental demands.

In general, the same pattern exists following the training session for the speaking dimensions assessed during the group presentation. Although reliabilities increased by 11.6% and 14.2% between the two checks for the content and language use dimensions respectively, they declined by nearly 8% for the application of the delivery or presentation style rubric. In sum, one quarter to over one third of the teacher raters had not accurately anchored into the four-point scales that were used for rating student performances as they left the training session.

Rater consistency was marginally better during the actual assessment in the classroom context, as Table 4 reveals. For the listening dimensions where an independent reviewer visited the classroom, 73.8 % of raters agreed with the judgment assigned by the reviewer when assessing student respect for peers and 70.5% for the participation dimension. But only two thirds (63.9%) agreed with the independent rating assigned for active listening skills. For the speaking skills rated during the group presentation, the consistency values were somewhat higher. Approximately four fifths of the teachers assigned an identical rating to that of the classroom visitor for the three dimensions of speaking competence assessed in the group communicative situation.

When comparing the results in Tables 3 and 4, it becomes apparent that rater agreement coefficients for the listening dimensions of communication remained relatively consistent, albeit low, between the training session and the actual judgmental exercise in the classroom. On the other hand, scoring consistency improved between the training session and the classroom assessment for speaking: when rating speech content teachers' degree of agreement with the independent reviewer increased by nearly 20%, and for presentational style the level of agreement improved by 15%. A smaller improvement (7.1%) was evident for the language usage dimension.

Although an independent reviewer was used for the classroom checks, all the variability cannot be ascribed to the teachers alone: the reviewer herself

Table 4
Rater Agreement[a] with Target Rating in Classroom Context, All Teachers, 1998 Saskatchewan Listening and Speaking Assessment (*N*=87)

Dimension	*Reliability*
Group Discussion (Listening)	
Participation	70.5
Listening	63.9
Respect for peers	73.8
Group Presentation (Speaking)	
Content	81.8
Language	81.8
Presentation style	78.8

[a] The coefficient represents the percentage of teachers whose rating on the 4-point scale was identical to that offered by an independent reviewer who simultaneously assigned a rating.

may not have demonstrated intrarater consistency with the two checks in the post-training session or between the training exercise and the actual classroom visitations. Nevertheless, it is evident that almost one fifth to as many as one third of the teacher-raters in this exercise were not calibrated with the six scoring benchmarks established in this exercise and as interpreted by the assessment coordinator. That scoring consistency improved for the speaking dimensions alternatively suggests that teachers were better able to assimilate the criteria during the interval between training and actual assessment; that they were able more consistently to judge live student performances than videotaped ones; or that a John Henry-like effect enabled teachers to become more consistent with an independent reviewer in the live classroom situation.

Reliability of the Assessment Design

A structurally sound assessment design should permit consistency of task situation from the point of view of participants. In Table 5 students' self-evaluation of performance and the students' evaluation of their peer group's performance are presented. For the group discussion elements six 4-point, Likert-type items were used, and for the group presentation elements three 4-point scales were used: this accounts for the different ranges in mean scores and their standard deviations. In terms of mean scores, students were slightly more stringent in assigning scores to themselves than they were in appraising their peers' performance in both discussion and presentation—which suggests stability in appraisal temporally and contextually across the domains. Moreover, the indices of variability in each domain for the self- and peer evaluations were virtually identical.

In terms of correlations, the highest (.63) was between self-appraised speaking skills through group presentation and peer-appraised skills in the group presentation. The lowest correlation (.44) was for peer evaluation of group presentation with self-evaluation of group discussion. For teacher ratings these dimensions are not comparable because the rating situations and instruments were different; but for this analysis, student ratings are comparable because they were completed post hoc at the same time with the same questionnaire instrument. The second highest correlation coefficient (.58) was for self- and

Table 5
Intercorrelations of Grade 8 Student Self Evaluation and Peer Evaluation of Group Performance, Group Discussion, and Group Presentation, 1998 Saskatchewan Listening and Speaking Assessment (*N*=589)

Variable	*1*	*2*	*3*	*4*
1. Student self-evaluation group discussion	1.00	.56	.58	.44
2. Student self-evaluation group presentation	.56	1.00	.45	.63
3. Peer evaluation group discussion	.58	.45	1.00	.58
4. Peer evaluation group presentation	.44	.63	.58	1.00
Mean	18.51	8.93	19.71	9.3
SD	3.03	1.90	2.98	1.87

Note. Group discussion rating is the sum of six, individual 4-point Likert scale items, whereas the group presentation rating is the sum of three, individual 4-point Likert scale items.

peer evaluation of performance in group discussion. These two highest correlations indicate some constancy of appraisal from within the domain of activity and by student perception from within the assessment situation.

Discussion and Potential Improvements

Reliability indices for the 1998 Provincial Learning Assessment in Listening and Speaking generally support the proposition that large numbers of teachers with training are able to synthesize both reasonably and consistently a wide variety of behavioral, audio, and visual information to render holistic judgments. Intercorrelations in the various dimensions of speaking and listening demonstrate that teachers can apply an integrated framework of appraisal to diverse groups of student performances in speech communication situations in ways that are internally consistent across geographically dispersed school contexts. Teachers were able to render holistic judgments using primary trait rating scales, in a manner that synthesized multiple oral language competences. The performance criteria in those scales encompassed a wide span of verbal and nonverbal behavior, which involved interpretations of individual and group intent and attitude and encompassed both audio and visual data.

Several aspects of the assessment and classroom practice may explain this ability to render holistic appraisals. Adjusting each of these practices, however, will not necessarily lead to greater internal consistency of the appraisal. First, the criterion rating scales were well developed and relatively precise in articulating the various attributes to be gauged. Second, the rating scales were supplemented by exemplar videotaped performances in training to render operational and integral the various dimensions assessed. Third, teachers may be accustomed to seeing students' speech performances in global terms in their day-to-day classroom work rather than as analytic subcomponents. Group dynamics (Battisch, Solomon, & Deluchi, 1993) and the ephemeral nature of speech communication may habitually demand of language arts teachers more molar judgments in the regular classroom context; in contrast, writing assignments with their tangible residue are more amenable to atomistic or analytic judgments that are customarily rendered with a pen. Fourth, by excluding students who were following modified programs, the actual sampling method may have yielded a narrower range of student abilities than generally is assessed; extremes of performance did not distort teachers' syncretic frameworks of appraisal. Fifth, the normative influence of group discussion may have had the effect of creating relatively homogeneous group performances, thereby narrowing the range of likely achievements. Sixth, the use of a four-point rating scale would sharply delimit the recognized range of performances discernible in the classroom. And seventh, the use of a single complex task to measure listening, followed by a single complex task to ascertain speaking proficiency, probably influenced the results because restricting the number of tasks has been shown significantly to reduce variability in listening and speaking tests (Brennan, Gao, & Colton, 1995; Brennan & Johnson, 1995).

This study generally supports the assertions of whole language teachers that holistic rating methods can be reliably applied in assessment situations beyond the written word. For the past two decades research into holistic scoring has repeatedly demonstrated its consistency in large-scale assessments of reading and writing (Hunter et al., 1996; Huot, 1990); the various analytic dimensions of literacy correlate well with holistically assigned scores. The strong scoring relationships between the various dimensions of speech communication suggest that with careful structuring of assessment tasks, listening and speaking need not be fragmented for reliable scoring of their multitudinous attributes.

However, assessors only modestly attained the goal of achieving rater consistency when applying these judgments in a decentralized assessment design. Teacher-rater constancy with predefined expert ratings at orientation sessions ranged from 51% to 74% on one occasion and 61% to 75% on another. In the school setting, interrater consistency among only two raters on a 4-point scale ranged from 64-82%. Between one fifth and one third of teacher scorers experienced difficulty during live marking in assigning ratings commensurate with the preestablished scoring benchmarks. Substantial percentages of teachers, particularly for the less discernible skills of group listening acuity, were not reliably "in tune" with the provincial "choirmaster," whatever the limitations of having this person's judgments as a criterion standard for validity purposes. Although rating consistency was better for the speech communication dimensions, the findings suggest that stability of measurement in the classroom may derive as much from external controls as from further internalization of the predefined standards through rater reflection in the period between training and actual live scoring (Bryant & Lee, 1995).

In reading and writing tests using criterion-referenced scales, reliability levels of .80 and higher are regularly reached with holistic rubrics (Huot, 1990; Williamson & Huot, 1993) in centralized scoring exercises for generating individual student results. That high percentages of teachers in this large-scale study of listening and speaking were not so consistent underlines the importance of rigorous preparatory training. If listening and speaking skills assessment is to be recognized by public authorities as trustworthy for comparing classroom or school-level performance, additional controls need to be brought to the design. As an alternative to single-teacher, decentralized scoring, students themselves could be systematically prepared with identical performance rating scales and exemplar video performances to assign rigorous peer ratings. In this way students' judgments could be combined with the teacher's to yield a multirater profile. Requiring teachers to conduct this orientation would enhance student expertise, should promote further assimilation of the criteria by the teacher himself or herself, and could engender greater consistency among teachers and students as a group—all with positive educational effect. Another approach would entail videotaping a sample of performances for scoring at a central location with multiple raters who have been systematically trained for removing individual biases and potential halo effects.

Neither improvement would automatically detract from the authentic assessment approach to speech communication except insofar as peers formally judging their counterparts' performance or the video camera are atypical of the classroom or social situation. Authentic assessment involves the administration of performance tasks and complex construction of responses in actual classrooms. We should not go so far as to say that evaluative ratings need to be assigned in the classroom context by the teacher alone as the singular audience. In other words, the 1998 Learning Assessment design for collecting information is not problematic; it is the evaluative procedures that require further adjustment to enhance reliability.

Indeed, the findings suggest that the administrative superstructure for this decentralized assessment was sound. Individual students and their peers were generally consistent in how they viewed their own and their peers' performances in a group context across the various dimensions of speech communication assessed. That they would do so indicates that viewed from within, the specific assessment situation and task were generally stable despite the diversity of classroom situations involved in the testing. Assigning the judgmental task through decentralization of scoring to classroom participants does not necessarily entail complete instability, as the classical measurement specialist might argue. The exoskeleton of centralized testing is not completely removed; rather, the internal regimen of a carefully prescribed task situation and the disciplined training of the scorer substitute for some of the more instructionally debilitating controls.

Conclusion

Given its centrality to school success, social status, and workplace effectiveness, we might anticipate that public officials should demonstrate concern with measuring oral communicative competence on a large-scale basis. After all, increasing demands for public accountability have spawned a number of large-scale testing programs ranging from the Council of Ministers of Education Canada's (CMECs) *School Achievement Indicators Program* (1999) through to the Ontario Educational and Accountability Office's initiatives. Yet few Canadian ministries have set out to assess and report on the oral and aural competences of public school students. Why is this so?

One reason is historic: the momentum of traditional achievement testing, the customary foci of public attention in language arts, and educators' previous professional preparation all predispose the education community to focus on literacy. Second, and perhaps most important, the time entailed in organizing teacher-scorers in a central location to listen to thousands of hours of audiotaped oral rendition when affixing marks is too financially daunting for a budget-conscious public administrator. The third reason revolves around the limitations of instrumentation: an audiotape directly captures only oral production, not listening acuity. Listening skill levels can be, and indeed traditionally have been, ascertained through multiple-choice test formats, but these types of instruments have difficulty capturing the interactive nature of listening in a social context, tending to measure hearing rather than listening. Conventional measurement devices have been unable to capture the dynamic nature of oral communication in a genuine classroom context. Videotaped performances can capture a number of dimensions, but may distort the performance because of examinee discomfort engendered by the camera eye.

As an initiative to investigate the speech communication skills of large populations of students in a socioeducational context, the 1998 Provincial Learning Assessment yielded mixed results. The global approach to judging various aspects of speech communication blending a variety of verbal and nonverbal traits shows promise. An assessment design that relaxes some external controls and substitutes a carefully prescribed task situation also appears viable as an assessment approach. Yet the decentralization of scoring was not successful in generating highly reliable data in this project. Assigning the judgmental tasks after preparatory training to teachers who are dispersed in classrooms across the province, rather than using carefully calibrated scoring teams at a central location, did not allow for stability in marking.

Although the achievement information was deemed sufficiently trustworthy for public reporting as a provincial-level indicator of the skill levels of Saskatchewan students (Saskatchewan Education, 1999a, 1999b), it would be unreliable for reporting classroom, school, or school division standings. Classroom practitioners can tolerate anomalies in the pursuit of better pedagogy, but measurement specialists would ask for a better balance of ecological validity with scoring reliability (Hambleton & Murphy, 1992; Moss, 1994) in a higher-stakes accountability situation. Further research into the scoring reliability of teachers with more intensive orientation and practice, with rubrics and video exemplars having greater trait specificity, and with parallel ratings from student peers may eventually correct this balance for both large-scale assessment and classroom evaluation.

References

Bateson, D. (1994). Psychometric and philosophic problems in "authentic" assessment: Performance tasks and portfolios. *Alberta Journal of Educational Research, 40,* 233-245.

Battisch, V., Solomon, D., & Deluchi, K. (1993). Interaction processes and student outcomes in cooperative learning groups. *Elementary School Journal, 94*(1) 19-32.

Brennan, R.L. (2000). (Mis)Conceptions about generalizability theory. *Educational Measurement: Issues and Practice, 19*(1), 5-10.

Brennan, R L., Gao, X., & Colton, D.A. (1995). Generalizability analyses of Work Keys Listening and Speaking Tests. *Educational and Psychological Measurement, 55*(2), 157-176.

Brennan, R.L., & Johnson, E.G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice, 14*(4), 9-12.

Bryant, C., & Lee, L. (1995). Group oral response to literature: An experiment in large-scale assessment. *English Quarterly, 23*(3-4), 15-22.

Council of Ministers of Education, Canada. (1999). *1999 School Achievement Indicators Program: Final report on reading and writing.* Toronto, ON: Author.

Dunbar, S.B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20,* 15-21.

Fredericksen, J.R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Hambleton, R.K., & Murphy, E. (1992). A psychometric perspective on authentic measurement. *Applied Measurement in Education, 5*(1), 1-16.

Hunter, D., Jones, R., & Randhawa, B.S. (1996). The use of holistic versus analytic scoring for large-scale assessment of writing. *Canadian Journal of Program Evaluation, 11*(2) 61-85.

Huot, B. (1990). Reliability, validity and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41*(2), 201-213.

McLean, L.D. (1990). Time to replace the classroom test with authentic measurement. *Alberta Journal of Educational Measurement, 26,* 78-84.

Mead, N. (1982). *Assessment of listening and speaking performance.* Paper presented at the National Symposium on Education Research, Washington, DC.

Morreale, S.P., & Backlund, P.M. (Eds.). (1996). *Large-scale assessment of oral communication: K-12 and higher education.* ERIC Clearinghouse on Reading, English, and Communication. Bloomington, IN: Speech Communication Association.

Moss, P. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5-12.

Nichols, P.D., & Smith, P.L. (1998). Contextualizing the interpretation of reliability data. *Educational Measurement: Issues and Practice, 17*(3), 24-37.

Quellmalz, E.S. (1991). Developing criteria for performance assessments: The missing link. *Applied Measurement in Education, 4,* 319-322.

Randhawa, B.S., & Hunter, D. (2001). The validity of performance assessment in mathematics for early adolescents. *Canadian Journal of Behavioral Science, 33,* 14-24.

Rubin, D.L. (1981). Using performance rating scales in large scale assessments of speaking proficiency. In R. Stiggins (Ed.), *Perspectives on oral communication assessment in the 80's* (pp. 47-52). Portland, OR: Northwest Regional Educational Laboratory.

Rubin, R.B. (1985). Validity of the communication competency assessment instrument. *Communication Monographs, 52,* 73-185.

Rubin, R.B., & Roberts, C.V. (1987). Comparative examination and analysis of three listening tests. *Communication Education, 36*(2), 142-153.

Saskatchewan Education. (1999a). *1998 provincial learning assessment in English language arts (listening and speaking).* Regina, SK: Author.

Saskatchewan Education. (1999b). *1999 Saskatchewan education indicators: Kindergarten to grade 12.* Regina, SK: Author.

Schippman, J., Prien, E., & Katz, J. (1990). Reliability and validity of in-basket performance measures. *Personnel Psychology, 43*(4), 837-859.

Stiggins, R.J. (1991). Facing the challenges of a new era in educational assessment. *Applied Measurement in Education, 4*(4), 263-273.

Stiggins, R.J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan, 77,* 238-245.

Swanson, D., Norman, G., & Linn, R. (1995). Performance-based assessment: Lessons from the health professions. *Educational Researcher, 24*(5), 5-11, 35.

Terwilliger, J. (1997). Semantics, psychometrics and assessment reform: A close look at "authentic" assessments. *Educational Researcher, 26*(8), 24-27.

Webb, N. (1995). Group collaboration in assessment: Multiple objectives, processes, and outcomes. *Educational Evaluation and Policy Analysis, 17*(2), 239-261.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70,* 703-713.

Wiggins, G. (1993a). *Assessing student performance.* San Francisco, CA: Jossey-Bass.

Wiggins, G. (1993b). Assessment: Authenticity, context and validity. *Phi Delta Kappan, 75,* 200-214.

Williamson, M.M., & Huot, B.A. (Eds.). (1993). *Validating holistic scoring for writing assessment.* Cresskill, NJ: Hampton Press.