Glossary of Assessment Terminology

Changes in the field of reading and writing assessment have generated a variety of new terms as well as new uses for established terms. The purpose of this glossary is to specify meanings for terms that are used frequently in discussions of literacy assessment.

Accountability

This term has dominated educational reform for at least the past decade. In its best sense, it means shared responsibility for constantly improving educational practices and short- and long-term educational consequences such as student learning and the qualities of the society the students develop. Policymakers, researchers, administrators, families, community members, teachers, and students all share this responsibility. Often, however, accountability focuses on the short-term responsibilities of teachers and students, such that primarily teachers and students experience the consequences when there are changes in achievement as measured by high-stakes tests. When teachers and students are held accountable only for short-term consequences, such as what can be measured on a test, longer term goals, particularly those not easily measured on a test, tend to be neglected. When only a subset of the community feels responsibility for educational improvement, education will not be well served and burn-out is likely to occur. An analogous situation would be holding a doctor accountable for a child's physical and mental health when the child has no health insurance (and therefore does not seek regular medical care) and his family's eating, exercising, and interaction patterns are not under the doctor's control.

Aggregation

In assessment, aggregation is the process of collecting data for the purpose of making a more general statement. For example, it is common practice for school districts to add together all students' test scores to find the average performance of students in the district. This process strips away all of the differences among the various cultural groups, schools, and students within the district in order to make the general statement. Even an individual student's test score is a result of aggregating all the items to which the student responded on the test to make a general statement about a student's "ability." It is also common to "disaggregate" scores to see how subgroups performed within the larger group or to investigate the students' performance in various subareas of reading (e.g., word identification, vocabulary, comprehension).

There are powerful tensions around aggregation reflecting, on the one hand, the need to make general statements about students, teachers, and schools and, on the other, the problem of stripping away the particulars of individual performances and situations in the process. Not everyone agrees that it is reasonable to reduce students or schools to numbers—let alone the purposes for or the grounds on which that might be done. It is often argued that administrators need highly aggregated data to make programmatic and budgetary decisions. However, both in education and in industry, administrators make different decisions when facing aggregated data than they do when presented with data about individual people and situations. Decision making needs to consider a balance of both kinds of data.

Authentic Assessment

For assessment to be considered authentic, it must include tasks that are a good reflection of the real-world activities of interest. This term arose from the realization that widely employed assessment tools generally have been poor reflections of what literate people actually do when they read, write, and speak. The logic of authentic assessment suggests, for example, that merely identifying grammatical elements or proofreading for potential flaws does not yield an acceptable measure of writing ability. Writing assessment tasks should reflect the audiences and purposes expected in life outside of school, with the real challenges those conditions impose. Similarly, reading very short passages and answering a limited number of multiple-choice questions is not a good measure of what literate people normally do when they read. Authentic assessments of reading employ tasks that reflect real-world reading practices and challenges. The authenticity of an assessment is very much a matter of the extent to which the assessment task measures what it purports to measure—a matter of construct validity.

Criterion-Referenced Assessment

We assess for particular purposes. When we want to know what children know and can do in a given domain, particularly whether they perform at a defined level on a specific task, we choose criterion-referenced assessment. Items in a criterion-referenced assessment are chosen because they discriminate what a person (or group) knows and can do and who has and has not reached a criterion level of performance. They are not chosen because they discriminate among individuals in determining who is better than whom. An item that genuinely measures a particular skill would not be eliminated from an assessment because everyone got it right. For example, a driver's test intends to determine whether a person is knowledgeable and capable enough to be allowed on the road, not whether one driver is more accomplished than another.

To be criterion referenced, a test must clearly define the characteristics that go into acceptable performance. In literacy, criterion-referenced assessments commonly compare students' performance on a specific task against established benchmarks. These benchmarks or criteria can be expressed as numerical ranges that define levels of achievement. For example, an 80–85 score may mean strong performance among levels of achievement ranging from unsatisfactory to outstanding. Criterion-based assessment can also involve holistic scoring of writing, for example, where a score is based on a set of pre-established criteria.

Compare to *norm-referenced assessment.*

Curriculum

We can think of curriculum as having three components: (1) the envisioned curriculum, (2) the enacted curriculum, and (3) the experienced curriculum. The envisioned curriculum is the intended proficiency of students as a consequence of instruction and participation in classroom events. The enacted curriculum is the daily attempt in classrooms to put the envisioned curriculum into practice. The experienced curriculum is the sense the learner makes of the enacted curriculum in the classroom and, thus, is constructed within the language of that classroom. For example, it is possible to intend to teach a particular lesson (e.g., authors' perspective) but that students not learn the lesson—either because it is not taught well (e.g., insufficient modeling, practice, support) or because the experiences of the students don't support the learning (e.g., they aren't provided with materials and experiences that invite perspective taking). As another example, if most of the reading material in one class includes racial or gender stereotypes, then that is likely to be reflected in students' learning. By contrast, students are likely to construct different knowledge about human relationships from a more balanced selection of reading material. However, the knowledge and attitudes students construct from those works are strongly influenced by the way teachers talk about them, the way teachers and other students respond to one another, and the nature of group discussions. Ultimately, it is the experienced curriculum that is our concern, and that is why students must be our primary curricular informants. However, the discrepancies among envisioned, enacted, and experienced curricula are what drive curriculum inquiry and the process of assessment.

Curriculum-Based Measurement (CBM)

This form of measurement was developed to help teachers evaluate a student's rate of growth in learning to read. The original idea was to have assessments that were embedded in the curriculum so they not only took no time away from teaching and learning but also did not distract teachers from the larger instructional picture. Originating in special education, a CBM of oral reading measures the number of words a child can read accurately in a minute from a standardized text (though there are comparable measures in spelling and writing). CBM assumes that a proxy variable, reading speed and accuracy (often mistakenly referred to as oral reading fluency), is an effective estimate of the larger construct of reading achievement and that the use of such estimates positively directs instruction.

Because these assessments now use texts and word lists that are standardized and that are not part of the curriculum, the term *curriculum based* is no longer particularly applicable. Other assessments not normally subsumed under the category of curriculum based, such as running records of children's reading and evidence of student work collected for a portfolio, are more clearly curriculum based since they are taken while the children are working within the actual classroom curriculum.

Equity

Issues of fairness surround literacy assessment. Testing originated as a means to control nepotism in job selection, providing an independent perspective on selection to uphold fairness. But equity cannot be assured through testing alone. Those who control the assessment process control what counts, what is valued. As we point out in this book's Introduction, language and literacy assessment is laden with cultural issues and biases. Although equity cannot be assured through assessment, it must be pursued relentlessly in assessment and in schooling. It is more likely to be achieved through the involvement of multiple, independent perspectives than through the use of a single perspective.

Tests have traditionally been administered, their results published, and their impact on instruction instigated with little regard to issues such as cultural, economic, or gender equity. But many equity issues affect assessment, rendering comparisons difficult and often invalid. Because traditional tests frequently reflect narrow cultural values, students and schools with different backgrounds and concerns often have not been fairly assessed.

Being equitable requires ensuring comparable educational experiences for those facing similar assessments, particularly in certification or gate-keeping situations. Questions of access to sound instruction, appropriate materials, and enriching learning opportunities are critical. Educators have become increasingly aware of the connections between assessment results and levels of safety, health, and welfare support in addition to physical accessibility.

Formative Assessment

Formative assessment, often referred to as assessment *for* learning, is the assessment that is done before and during teaching to inform instruction. It is assessment that informs instruction. Formative assessment includes things like teacher–student conferences, listening in on student book discussions, taking records of children's oral reading, examining students' writing pieces, and so forth. Though these assessments might be standardized, they often are not. To be formative, an assessment must affect instruction.

Compare to *summative assessment.*

High-Stakes Testing

These tests have significant consequences for those viewed as responsible for performance on the tests, and also for the student. For example, tests that determine whether one is accepted or rejected into the military, a university, or an educational program have significant consequences for the individual test takers. Consequences can be felt among a broader range of people, however. In the United States today, student test scores are not only used to determine whether children move on to the next grade level, but they also influence where educational resources are allocated and whether a school may continue to operate. Often, local news media publish school test scores, and property values are affected when families make decisions about where to purchase a home based on the local school's performance. When major consequences—such as the adjustment of teachers' salaries—are attached to their students' test scores, teachers will emphasize in their instruction what the test measures and reduce their emphasis on areas not covered by the test. This has consequences for the breadth of the curriculum and, thus, for the students' lives.

Both the National Council of Teachers of English and the International Reading Association have position statements regarding high-stakes testing. Both organizations recommend minimizing the stakes where possible and not relying on single measures, particularly when the stakes are high.

Inquiry

The process of inquiry begins with a genuine question, that is, a question that motivates the questioner to persist in seeking the answers. Authentic questions are rarely well formulated or structured at the outset. Rather, structure emerges through the process of inquiry. Inquiry is not merely a matter of asking and answering questions. It is a way of engaging the world and other people. Communication and social relationships play an important role in inquiry as questioners seek the advice and expertise of peers and more knowledgeable others, share their findings, reflect upon the results of the inquiry, and take up new questions that arise.

In a traditional view of classroom learning, teachers deliver information. They ask the children questions to which they already know the answers, and the students are to show they know the correct answers as well. This approach has not been very successful at helping all students become the critical, creative, and socially responsible citizens our society needs. In an inquiry classroom, on the other hand, students and teachers have a different relationship. Teacher and peers are resources for helping students answer their own questions. The community relationships are different. Instruction is based on engaging in sustained examination of personally significant topics.

Assessment as inquiry involves the same principles. It requires teachers to pose questions about the teaching and learning in their classrooms and to seek answers to those questions using assessment data and the resources of their learning community.

Multimodal Literacy

For centuries, the book has been the central medium of communication, expressed on paper largely through the mode of writing. Today, the screen is becoming the dominant medium of communication, with increasing reliance on the mode of image. A mode is a resource for communication and representation. Examples include speech, dance, gesture, music, sculpture, photography, and writing. Humans may express themselves through a single mode, such as writing, but with growing frequency we combine modes to communicate. This results in multimodal texts such as a PowerPoint presentation or YouTube video that combines words, images, music, and movement, or an advertisement in which print and image are merged. Today's and tomorrow's learners need to acquire competence in this multimodal literacy.

Norm-Referenced Assessment

When we want to know how a child performs relative to other children in a particular domain, we use norm-referenced assessment. Items in a norm-referenced assessment are chosen because they discriminate between individuals rather than assessing what a person (or group) knows and can do. To make norm-referenced assessments, assessment practices need to be standardized and test item selection must focus on maximizing the differences among individuals on a scale. An item that genuinely measured a particular skill but which all students got correct would not be used because it would not discriminate who was better than whom.

Norm-referenced interpretations are based on comparisons with others, usually resulting in a ranking. For example, a norm-referenced interpretation of a student's writing might assert that the sample is "as good as that of 20% of the students in that grade nationally."

Norm-referenced testing is the most prevalent form of large-scale testing, in which large groups of students take a test and the scores are grouped and interpreted in relation to other scores. In other words, the score of any student or group (school, district, state, or nation) has meaning only in relation to all the other scores of like entities (e.g., school to school, district to district, state to state). In order to make such comparisons, we have to make the assumption of "all else being equal," which is rarely justifiable. National norm-referenced tests assume that all students in our society have had similar cultural and curricular experiences. Uses of these tests also commonly ignore differences in curriculum, culture, gender, ethnicity, economic circumstance, per-pupil funding, and so forth.

The main advantage of such assessments is the simplicity of the linear scale. The seductiveness of this scale is also the main disadvantage, because the scores appear readily interpretable and objective. However, the score oversimplifies the complexities of literacy and assessment. Unfortunately, norm-referenced test scores often become the most important criterion for decisions about placement and promotion, which have a powerful impact on students' and teachers' lives.

Compare to *criterion-referenced assessment.*

Performance-Based Assessment

Performance-based assessment refers to assessment that involves the demonstration of a particular skill and often the process of accomplishing a performance specific to that skill. Performance assessments can include, for example, such complex activities as group collaboration to write and produce a play. The concept of performance-based assessment is related to the concept of authentic assessment in that it arose from a realization of the limitations of multiple-choice tests, and other assessments of complex skills, and the difficulty in making inferences about complex skills from such assessments.

Portfolio Assessment

A portfolio approach to assessment uses a systematic and multifaceted collection of work that represents a student's development. For example, a portfolio might include a range of writing pieces, a book log, self-reflections, group projects, and multimedia work. Because of the nature of the contents, portfolios are both curriculum based and performance based. A primary emphasis in most portfolio assessment is on student involvement and the development of self-assessment or reflectiveness. However, in some applications, portfolios can also include teacher and parent observations.

Reliability

Broadly speaking, reliability is an index of the extent to which a set of results or interpretations can be generalized over time, across tasks, and among interpreters. In other words, it is a particular kind of generalizability. For example, a common concern raised by newer forms of literacy assessment is whether different examiners, evaluating a complex response and using complex scoring criteria, will draw similar conclusions about a student's performance (whether an assessment will generalize across different examiners). Experience from scoring complex student writing samples suggests that high rates of agreement can be achieved when people are well trained in the application of specific criteria.

Another example of reliability is whether a score obtained by a student on a test would remain the same if the student took the test the following day, assuming no new learning has taken place—in other words, whether the performance generalizes over time. In general, the more samples of student work we collect, the more reliable and consistent an assessment will be.

Reliability is only important within the context of validity—the extent to which the assessment measures what it is supposed to measure and leads to useful, meaningful conclusions and consequences. Reliability does not guarantee a high-quality assessment. It is possible that consistent scoring can be achieved on poorly designed tests or tests of trivial skills. Indeed, reliability is easiest to obtain on low-level skills.

Summative Assessment

Summative assessment, often referred to as assessment *of* learning, is the after-the-fact assessment in which we look back at what students have learned, such as end-of-course or end-of-year examinations. The most familiar forms are the end-of-year standardized tests, though in classrooms we also assess students' learning at the end of a unit. These assessments are likely to be uniform or standardized.

Compare to *formative assessment*.

Validity

Historically, a common definition of a valid measure is that it measures the construct it purports to measure. This is called *construct validity*. For example, if we claim that an assessment measures reading fluency, but it only measures speed and accuracy and does not include aspects such as intonation, the test would have poor construct validity.

More recent conceptions of validity include an examination of the consequences of assessment practices—*consequential validity*. For instance, a test might have excellent construct validity as a measure of decoding ability. However, if it were used as the basis for adjusting teachers' salaries, resulting in an overemphasis on decoding in the curriculum, it would not be a valid assessment process. In other words, one cannot have a valid assessment procedure that has negative or misguided consequences for children. Consequently, a productive definition of a valid assessment practice would be one that reflects and supports the valued curriculum.