Wave of New Assessment Paradigms

The assessment of reading and literacy performance has been regarded a key proxy for judging advances and improvements in education. Those who have aspired to make education a science have viewed reading achievement as a highly valued and readily measurable outcome that could be indexed by simple observable behaviors, such as oral reading accuracy and fluency or scores on a multiple choice test of passage comprehension (Pearson & Hamm, 2005; Resnick & Resnick, 1988, Resnick & Resnick, 1992). These yielded scores by which readers and schooling could be compared for purposes of policy considerations, research or educational planning.

Over time, since their appearance about the time of WWI, tests have assumed more and more prominence. Testing became integral to making educational policy decisions as well as aids to instructional decision-making by teachers, administrators as well as the public at large. They were enlisted by universities for admissions screenings, at times in efforts to limit and discriminate against certain students and at other times prompted by ideals associated with meritocracy (Lemann, 1999; Willis, 2008) . By World War II, tests were extensively used to assess the preparedness for recruits for different posts. And, after World War II in the era of the Space Race, the media often looked to test results to portray the success or failure of schooling in terms of alleged shifts in performance. In a number of ways, tests became the gatekeepers for societies and in turn had a major impact upon schooling. For many students and societies, tests served as lifelines to opportunities; for others, as screening filters to track, label or prescribe. Tests provide for the assessment of groups and the differential assessment of individuals, but the results of tests can also be used as gatekeepers. As the stakes, for both schools and individuals, became higher and higher, tests became a determiner of what was taught and emphasized—a kind of default curriculum.

Standards for validity, reliability, fairness, and utiity are intended to guide test construction and use. Notions of validity (content, construct, concurrent, predictive and consequential) are tied to concerns about: 1) Whether an assessment represents the *content* (topics, processes, skills, strategies and outcomes) of what should be measured; 2) Whether the test correlates well with other assessments of the construct (that's concurrent) or is a good *predictor* of performance on some future criterion (e.g., alphabet knowledge at kindergarten predicting reading achievement at the end of grade 2); and 3) Whether the test themselves and results have *consequence*s that are beneficial or problematic to those who are assessed. Checks on the reliability of assessment instruments (test-retest, intra-test reliability) are enlisted to check on issues of consistency and the power of the tests to offer stable measures of differences between groups or individuals. In particular, test-retest reliability addresses whether performance on a test would be relatively consistent if the test was readministered; internal consistency reliability assesses whether items measures the same construct consistently. Depending on the test's degree of reliability, tests might be reasonable for group use or have sufficient precision to offer safe judgements on or between individuals.

In terms of matters of interpretation, assessments are often calibrated against normative data for students from appropriate comparison groups, such as similar age groups, as well as the skills and abilities represented by the performance of the test (e.g., the strengths, weaknesses, mastery or other measures of performance). And, as test use has increased, the tests themselves have become streamlined, responses digitized, and scoring increasingly automated, the results are available expeditiously—sometimes almost immediately.

Reading tests have almost become a genre unto themselves with similar features to one another. Reading comprehension measures are predominately based upon a reader's responses to questions following the reading of a paragraph or extended text; oral reading accuracy is still based upon a form of read aloud of selected passages of increasing difficulty. Changes occurred with the addition of a few dimensions informed by research and some shifts in the reading curriculum implemented in schools. For example, with vocabulary emerging as a key predictor of reading comprehension (and appearing seemingly easy to assess), vocabulary subtests became an addition to most reading assessments. In the 1930s, with the advent of the notion of reading readiness coupled with observational studies of reading development, assessments of young students expanded to include measures deemed to be predictors of early reading, such as visual and sudatory discrimination, and letter name knowledge.

Also commonplace were forms of testing that dovetailed with developments in readability, as well as views of reading progression aligned with a linear demarcation of the difficulty level assigned to reading material. The introduction of readability formulae fueled the view that the difficulty level of reading material could be reduced to a formula, based upon assessment of the vocabulary that was used and the complexity of sentences (usually measured by length). Once reading material was assessed as corresponding to certain grade levels, educators developed procedures to place and track the development of readers. Among the most common was a subjective evaluation of reading comprehension and oral reading labelled as the Informal Reading Inventory. The key criteria for placement dated back to what was referred to as a five finger rule, where placement in reading material was deemed to correspond to 95% oral reading accuracy (i.e., only 5 errors in a 100 word selection) and 75% comprehension accuracy. Nowadays the practice continues in variations of this same form of levelling and assessment. For example, in programs such as Reading Recovery, running records are used repeatedly to assess the progress and decide on the advancement of the students to more difficult reading material.

Consistent with the interest in readability, the cloze technique (with ties to journalism) became widely enlisted as a way to judge comprehension (Taylor, 1953). As the name implies, the cloze procedure required the reader to complete passages—usually a few hundred words in length—where words were deleted systematically throughout the passage (typically every fifth word was deleted with the first and last sentence left intact). On the assumption that the reader's accuracy to complete the passage corresponded with the reader's comfort reading the material, it was then discerned whether or not the material was within the comfort zone or beyond the reach of the student. While systematic studies of cloze suggested that it was not without flaws, it became widely used as a quick guide for checking on whether a passage was at a suitable level for a reader (e.g., Shanahan, Kamil & Tobin, 1982).

The massive growth of various forms of tests of reading or reading-related skills, as well as protocols for testing, contributed to extensive reviews of the available tests and their properties. Indeed, ongoing testing became integral to the curriculum developments of the 1960s and early 1970s, in part as a systems approach was increasingly enlisted to monitor students' mastery of what were deemed the skills of reading. This included what is regarded as a criterion-based approach, in line with systems analyses.

Nowadays, various forms of tests are available with most curriculum or as distinct tests for particular needs. The quantity and variety of tests is overwhelming; reviews of thousands of published educational tests are compiled in various volumes of the Buros Yearbook (published by The Buros Center for Testing, a non-profit located at the University of Nebraska-Lincoln). Tests to measure reading performance still proliferate, especially with ongoing government mandates tied to accountability. There remain periodically reviews of tests conducted by literacy educators (e.g., such as that advanced by Roger Farr) and testing has been the subject of critical reviews in most major reference works in literacy. This is especially evident in selected books or articles focused on different forms of evaluation, including comprehension assessment (Johnston, 1983), portfolio assessment (e.g. Tierney, Clark, Fenner, Herter, Simpson, & Wiser, 1998), or major syntheses critiquing and tracing the history of reading assessments and their relationships to learning and policy (Johnston, 1984; Calfee and Hiebert, 1991; Valencia & Wixson, 2000).

A map of the terrain of testing is now therefore quite multifaceted and somewhat multidirectional as a number of different assessment pathways are pursued. These include forms of:

1. External testing for the following goals:
   - Large scale standardized or comparative assessments to monitor educational progress on the international, national, state and regional level over time and across jurisdictions and populations, e.g. countries, states, urban, rural, ethnicities, language groups, gender, and backgrounds (e.g., Global, PISA; United States, The National Assessment of Educational Progress-NAEP);
   - Admissions testing and minimal standards testing for purposes of judging qualifications tied to tracking, acceptance, promotion and graduations (e.g., Graduate Records Examination; TESOL, SAT);
   - Screening devices for special services, such as special needs (e.g., Woodcock, ITPA);
   - Periodical assessments of schools, teachers, classrooms and students befitting the demand for local accountability tied to expectations for progress;
2. Internal testing for purposes of:
   - Teacher assessments of students to guide instruction, including:
     - Criterion-based measures tied to a prescribed set of skills aligned with what has become labelled RTI (Response to Instruction);
     - Periodic informal assessments of the student's overall reading ability, including the level of reading and a profile of developments derived from a mix of ongoing teacher observations, checks, rubrics, etc. that might be related to school work and assessments emanating from running records (tied to occasional checks on reading accuracy, understandings, and strategies);
   - Student-based assessments, directed at helping students deliberate on their own efforts, processes, pursuits and achievements tied to projects, portfolios and rubric discussions.

Some Critical Perspectives on Reading Assessments

Testing became a dimension of society that in many countries, such as China, overshadowed what was emphasized in schools. It has been a major part of the landscape of education and has remained so despite some efforts to shift the forms and uses of tests as well as the epistemologies that undergird them. The use of testing should not be considered uncritically. Historically, assessments claim to advance meritocracy; in reality they may have served nefarious purposes. Indeed, reading assignments were used as a means of screening persons for positions and eligibility to vote or to define their eligibility to access opportunities, both economic and social. Indeed, the entrance examinations for universities were initially intended to serve as a means of excluding certain groups or reproducing social privilege and class control of education. In the interest of uniformity, tests have served as ways to impose cultural norms upon others in ways that have displaced cultural ways of knowing.

While tests were touted as fair, they were not constructed to measure the diverse literacies of test-takers—treating representativeness as secondary to uniformity. Most standardized tests had an inherent bias as a result of their efforts to ensure uniformity; a bias to mainstream communities. In particular, those involved in major test development have tended to exclude passages or items where student responses vary. Unfortunately, in the interest of aggregating scores and pursing matrix sampling, a passage may be excluded if students' performances seem erratic (despite being perhaps consistent with their background knowledge). In reality, readers will perform quite differently from one text to another, depending upon the relevance of what they are reading. Depending upon the topics chosen, any one person or group may be advantaged over another. Yet often test makers proceed almost with sleight of hand as they fit tests into their psychometric models—with preset biases to the view that reading ability should be homogenous rather than varied (see Side Comment III.6b.1).

Side Comment III.6b.1.

It is troubling that nowadays many large-scale tests seem to be developed with assumptions which are not aligned with the complex nature of reading. In particular, what is troubling is that they can use test discrimination coefficients to weed items without regard for relevancy or representativeness.

Of course, you can ignore these complexities and assume that tests represent a genre that has become acceptable as a proxy for real reading, or present a reasonable measure of reading. Yet if the goal is to achieve a truer measure of capabilities, then you are faced with the dilemma that your results may well predict the outcomes on other similar tests, but not to reading in the real world—especially for non-mainstream students.

Looked at historically, tests quickly became entrenched in what was measured, how, when and why. It is as if they became "the tail that wagged the dog" as curriculum became tied to the test and not the reverse. Despite calls to update curriculum, the test dictated the path forward. Indeed, calls for revolutionary changes in curriculum were usually nullified as a result of the testing traditions that had taken hold—especially if the tests were high stakes (e.g. tied to graduation or access to educational opportunities such as university entrance). Teachers and students, in turn, would invest in learning for the test. What was taught was aligned with what and how reading was tested—typically multiple-choice responses to short passages.

For example, take the forms of reading that constitute many national and international tests and most standardized tests. Tests are mostly built upon a sampling of texts that may or may not match the test takers' experiences or pursuits. Multiple choice items measure a person's response to having to make a choice in accordance with the set-up of the questions and options from which they will choose. If tests employ questions with open-ended responses, how the questions are asked and support a response will have an influence on what a test-taker provides. In tests with retellings or recall, there are differences in how comfortable readers feel when asked to freely retell the text, and different degrees of responsiveness to further probes. A test's measure of reading is more a measure of what and how the test is measuring.

Indeed, if you compare reading and writing in the real world with reading and writing in the test world, there are many differences:

- Whereas in the real world, individuals encounter a wide range of different materials enlisted for a range of purposes that have variable relevance, most tests are limited to a small subset of passages.

- Whereas in the real world, individuals live in the media as they engage in array of connected digital texts and environments (e.g., emails, text messages, the internet, social networks, etc.), tests tend to use print forms or represent online copies in print forms.
- Whereas in their real worlds, individuals are engaged in texts connected to their worlds—that is, those that are culturally-rich and apt to vary by gender, age, and culture—traditional testing vies to be culturally free (that is, homogenized or standardized).
- Whereas in the real world, reading is often pursued with others—in a fashion that is collaborative—tests involve reading by oneself.
- Whereas in the real world there is more of an acceptance and recognition of different interpretations and views of texts, in tested reading there is an assumption of correct responses only.
- Whereas in the real world reading may involve various forms of extended, incidental, and impromptu reading events, integrated into ongoing literacy engagements, tests involve a limited form of reading—one that is scheduled and often laden with time constraints. Similarly, whereas real reading involves a range of purposes, from incidental to emerging to relevant, tests prescribe reading purposes.
- Whereas in the real world, we approach texts with different intentions, intensities, and approaches (tied to access and uses, including ongoing inquiry, keeping us up-to-date, or satisfying our own indulgences), tests involve responding to selected, pre-set test formats—to be responded to under formalized test conditions and durations.
- Whereas in the real world, our engagement with text is more likely to represent a mix of communications with others (i.e., ongoing projects, articles of interest, memos, emails, text messages, and web searches—akin to a messy desktop with numerous files, images, tags, and invites for comments or feedback), our test engagements represent a restricted array of material and probes, with limited connections to our lives, including our next steps or ongoing decision-making.
- Whereas our reading in the real world is tied to our actual pursuits, tests are tied to attempts to yield scores or evaluation summaries that assess and compare our reading and try to predict to the real world.

The nature of reading development adds to these complexities. Test-makers and curriculum developers perpetuate a false illusion that there is a set sequence to reading development; that a single score can be ascribed to the stage of a reader's development. But learning does not progress in a fashion that is unidimensional or linear. Development profiles of readers may be "shoe-horned" to fit a preset, simplified scoring system, but a truer representation would be multifaceted, multidimensional and varied in accordance with differences between readers. If you consider your own reading development, you would probably describe your reading performance as varied—dependent upon, for instance, whether your reading was in areas for which you have an expertise or experience or in areas for which your experience is less informed. Your literacies are somewhat unique or not as not uniform as what and how it is apt to be measured. You may read a lot of political discussions, especially those pertaining to foreign relations and to certain politicians with whom you are intrigued. Your reading may be tied to certain authors or topics from sports to human interest to self-help matters. If you were to be profiled, it would be somewhat unique and likely to change over time. It would look different when compared with the profiles and developments others and not follow a prescribed sequenced (See Side Comment III.6b.2).

Side Comment III.6b.2

We should therefore be careful not to overgeneralize the merits of using piecemeal, step-by-step learning progressions, which might befit the learning of a narrow set of skills and understandings (e.g., learning how to construct a PowerPoint or do certain mathematical processes). Some of my colleagues would suggest that over time we may have the research that would enable us to generate a development trajectory for which we could confidently predict individual reading outcomes and development (e.g., Shepard, Hannaway, & Baker, 2009). But too often, educators turn tests into instructional regimens, wherein test specifications (especially those descriptions of the characteristics or signs of reading development that undergird test designs) shift into prescriptions for development, the bases for curriculum, or guides to teaching.

Stepping Back

So, what are the ramifications of measuring reading?

- The reality of reading is that if your goal is to attain a true measure, then your approach needs to be robust and likely diverse, as the nature of meaning making is not standardized and varies across readers and across circumstances for reading.

- Your measures should allow for variability across readers and acknowledge the fallibilities of tests.
- If you were to describe reading development over time, you would need to adopt means for doing so that reflect the nature of varied development in ways that capture the complexity (rather than distort it).

Why might you choose to accept the current testing regimen?

- Tests in their present form are accepted, respected, and expected in society.
- Test developers are seeking more straightforward forms of comparison and means of aggregating individual and group performances.
- Tests are viewed as adequate proxies for more complex profiles.
- Test scores can be easily normalized, and are useful for educational decision-making and, over time, for student learning (e.g., as checks on responses to teaching).
- You can use broad enough developmental categories and adopt statistical procedures (e.g., Item Response Theory, IRT, tied to uni-dimensional modeling assumptions) to force-fit or selectively pull together a set of items to conform to such a model.
- Summative test results provide help to policy makers as they make decisions.

Why might you contest the dominant regimen of most tests?

- Tests may sample texts, but their sampling represents a limited array of the types of texts encountered in the real world.
- Tests represent an interrogation of readers that is a step removed from reading in the real world.
- Tests may stagnate and narrow the curriculum.
- Test scores are not a proxy, nor are the scores useful for educational decision-making or student learning, as they do not adequately represent reading in the real world.
- Summative test results may not help teachers teach and students learn.
- Test scores are tied to notions of reading development and ability as uni-dimensional, when reading ability is not uni-dimensional.
- Tests often marginalize minorities in what and how they test—sometimes keeping individuals and groups invisible as tests are developed and results are reported.

Most tests may give the illusion that they are measuring reading, but they may just be measuring themselves. Unfortunately, the illusion may be supported when we teach to the test, a practice which is sometimes rampant. Tests seen as representative of a form of aberrant reading, when elevated in status, do lead to a coercive and potentially erosive influence on reading development. Indeed, a test should be viewed systematically; it should measure up to ethical standards tied to judicial decision-making and responsive evaluation (Guba & Lincoln, 1989; Lather, 1986; Moss, 1996) in terms of its design but also in terms of how it is used. Developers and consumers of tests should view as integral to their use a consideration of the consequential validity of their measures—including the impact of tests upon students, teachers, parents, communities and society. As Kris Gutiérrez (2004) has argued, incidences of teaching to the test involve:

> …a set of complex issues that defines schooling for so many students today…. It is an account of the consequences of narrow views of literacy and how a teacher's understanding of literacy is complicated and constrained by a mandated school curriculum that was conceptualized and implemented independent of the knowledge and practices of its students. It is an account of the ways we understand competence across racial, ethnic and class lines. It is an account of the consequences of the ways we measure what counts as literacy, especially if we only see it in snapshots in discrete moments in time disconnected from the laminated, multimodal reality of literacy activity. (p. 102)

Making Progress by Changing the Underlying Tenets of Assessment

As constructivist and socio-cultural models of literacy grew in prominence and critical perspectives achieved traction, a number of educators in the 1980s and1980s turned their attention to the nature and role of assessment. Examined from the perspectives of policy personnel, teachers and learners, questions were asked about how well assessments served the needs of teachers and learners in advancing in all their diversity at the classroom level (see Side Comment III.6b.3). Likewise, literacy educators recognized that literacy assessments were dated and did not match how literacy was currently viewed, in terms of the tenets of constructivist research or the growing emphasis of teachers as professionals and students as strategic learners (e.g., Valencia & Pearson, 1987). These criticisms included: 1) Concern that multiple choice tests or other forms of "closed" assessments did not align with constructivist tenets; 2) The emphasis upon tests was detracting from efforts to advance classroom practices; and 3) If classroom practices advanced, they would not match with testing practices.

Side Comment III.6b.3.

As the U.S. Congressional Office of Technology Assessment (1992) stated in the summary of a report, Testing in American Schools: Asking the Right Questions*: "The move toward new methods of testing has been motivated by new understandings of how students learn as well as changing views of curriculum..." (p.16). They argued for forms of performance-based assessment that better fit with learning that was relevant and meaningful, intending for tests themselves to serve educative as well as evaluative functions.*

In response, education witnessed a surge in what was termed "authentic assessment' practices, including a range of tools to support their use. Project-based learning assignments with assessment components proliferated. These were accompanied by a range of teacher adjuncts, such as the use of dynamic forms of record keeping including anecdotal records (Barr, Ellis, Hester, Thomas, 1988), running records (Clay, 1993), and retellings (Irwin & Mitchell, 1983; Morrow, 1988). Also notable was the rise in the use of portfolios among educators.

These new approaches represented a form of engagement with students that dovetailed with classroom pursuits, especially forms of project-based learning. They correlated with the increased emphasis upon reading-writing connections and student-based conferencing with student self-assessment. A number of educators saw the advantages of such measures over other modes of assessment, especially in terms of better representing student literacy development and processes in more complex ways than scoring systems. Studies suggested that parents were also more apt to prefer these approaches as a better representation of a student's learning (e.g., Shepherd & Bleim, 1995). Studies comparing forms of performance assessment with traditional measures suggest that changes in learning are more apt to be captured by such assessments (see Shavelson, Baxter & Pine, 1992; Tierney, Clark, Fenner, Herter, Simpson, & Wiser, 1998).

The Reform Movement Agenda and Setback

Despite the momentum of these developments in the 1990s, traditional approaches to assessment regained their prominence with the rise in accountability tied to school reform efforts. These were anchored in large scale assessments within and across countries for students, schools and populations. Indeed, policy makers seemed intent on using reports from traditional assessments to judge and plan educational progress, providing report cards of progress based upon measurable outputs rather than on more diverse qualitative considerations including inputs. Despite the recent U.S. NAEP team's proposed approach, socio-cultural considerations have had a tendency to be sidelined or displaced by reports of the performances of different groups by race, gender, and location. In reality, these reports, which herald, applaud, or deplore the performance of different groups, seem to ignore some of the history of testing—especially its use to exclude as well as its inherent bias tied toward the erroneous notion that such tests are and should be culture-free (see Side Comment III.6b.4).

Side Comment III.6b.4.

For example, some of the current U.S. national assessments provide pull-down, digitally-based menus, the results of which can be examined according to the ethnic makeup and location of test takers. However, these sites have failed to include items that are responsive to the literacies of different groups as well as those that are up to date with the media world we inhabit. Test developers seem to have had an appetite to apply advances in technology to test reporting, but have not as yet embraced these technologies as a means of supporting different, more representative sampling across various groups.

Tests have a history that seems almost chameleon in nature—touting to be something other than what they are, but constrained by the framework from which they are derived. Certainly, the history of reading suggests that the testing of reading has been limited to the tenets of what was observable and score-able; hence, simple additive dimensions—such as the accuracy of responses to questions or vocabulary assessment—were enlisted to measure ability rather than more complex and idiosyncratic engagement. Reading tests have a history of using summative scores of ability or target scores for mastery that are quite contrived. Unfortunately, they often dictate decision-making, as if they are more trustworthy and less limited or restricted than they are. And, unfortunately, this seems to have been advanced with the marriage of test scores to standards and curriculum, as endeavors such as Response to Intervention (RTI) (Fuchs & Fuchs, 2006) and the widespread use of tests such as DIBELS (Good et al., 2001; Riedel & Samuels, 2007) have been viewed as positive, without being questioned (Goodman, 2006).

In Closing

As a number of scholars of assessment[1] and others have suggested, test developers and users should be held to standards related to a test's validity and reliability, but also to ethics—considering a test's consequences, including the social and educational ramifications (e.g., Calfee and Hiebert, 1991; Johnston, 1984). As Bruner (1990) argued, we should be "conscious of how we come to our knowledge and as conscious as we can be about the values that lead us to our perspectives. It asks us to be accountable for how and what we know" (p. 30). Examined historically, testing has had a questionable history. Literacy assessments especially have a history of a role in society that may be nefarious (Ellwein, Glass, Smith, 1988; Haney, 1993; Nichols & Berliner, 2007; Willis, 2008). Quite deliberately, literacy tests have excluded individuals and groups (i.e., in terms of access to voting, education, or other rights). Indeed, under the guise of being a reliable and fair test for all, tests often reflect a questionable alignment that advantages some over others. If they assume or are given a high stakes status, there is a danger of being extremely harmful. As Campbell (1979) suggested:

> The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt is will be to distort and corrupt the social processes it is intended to monitor. (p. 85)

Yet many have argued that the solution lies with improved forms of testing, and in recent years we have seen some inroads again. In an effort to pursue major improvements in the assessment of reading comprehension, the National Academy of Education released an expanded discussion of *Reading for Understanding (RfU)* that included a lengthy critique of reading comprehension assessment and the pursuit of what they suggested were "forward-thinking assessments that not only meet the standards of educational and psychological testing, but also promise to advance both research and practice in reading comprehension for years to come (Pearson, Palincsar, & Biancarosa, 2020, p. 255). Among those assessments touted in the RfU report was what has been identified as *scenario-based assessment*, incorporating a project-based frame befitting a more realistic form of reading comprehension assessment. In concert with a shift to what is learned rather than comprehended, this enlists

[1] Including Gordon Stanley, Lee Cronbach, Gene Glass, and, more recently, David Berliner, Robert Linn, Lorrie Shepard, Ernest House, Pamela Moss, and Sean Reardon.

project-based stimuli and forms of outcomes befitting the learning goals (O'Reilly, Sabatini, & Wang, 2018; O'Reilly, Weeks, Sabatini, Halderman, & Steinberg, 2014; Sabatini, O'Reilly, Halderman, & Bruce, 2014a; Sabatini, O'Reilly, Halderman, & Bruce, 2014b; Sabatini, O'Reilly, Weeks, & Steinberg, 2016; Sabatini, O'Reilly, Weeks, & Wang, 2019).

Consistent with the sentiment for a new framework to guide literacy assessments, the United States design team for the 2025 National Assessment of Educational Progress in Reading (NAEP) have also shifted to a form of testing that is more governed by socio-cultural tenets. As the design team suggests, the key components of the (ref: sociocultural) "…model—reader, text, and activity—are situated in both highly specific contexts, such as classrooms, homes, or digital spaces, and more general contexts, like communities, social networks, and nations" (NAEP, 2020, p 19). They do so by investing in an approach to passage and task selection that is more representative of the diversity of texts and situations experienced in the real world. Enlisting a form of scenario-based testing, assessments are scaffolded (including with avatars) in an effort to mimic "real world reading." As they state:

> The most fundamental principle of the sociocultural model of reading is that, as a human meaning-making activity, reading is always situated in social and cultural contexts that shape every aspect of readers' engagement with text and influence how readers respond to and learn from the experience of reading. The Assessment Construct reflects this understanding by using testing blocks that are highly contextualized. NAEP 2025 assumes and attempts to build on the cultural assets (knowledge, skills, and practices) that all students bring to the assessment. (p. 28)

To this end, the NAEP team is pursuing an emphasis focused on mimicking situated or contextualized literacy tasks. As the team states:

> This emphasis on contextualization is present from the moment readers begin the NAEP 2025 assessment. For example, at the outset of an assessment activity, readers will be introduced to what will be called an activity structure. That introduction will specify a simulated context for traversing an entire 30 minute block, including:
>
> - A simulated social setting (a community setting or a classroom and perhaps some avatar classmates and even a teacher) and an explicit role for the reader

- A purpose for engaging in the entire activity (an activity-specific instantiation of one of the two overarching Purposes (Reading to Develop Understanding or Reading to Solve Problems)
- The disciplinary Context in which the activity is situated (Literary Context, Science Context, or Social Studies Context) (p. 29)

The pursuit of such a shift represents a major development consistent with the advances in literacy. The effort to simulate is admirable but immensely challenging, especially if outcomes are to be reported in a fashion befitting the tenets of the effort. However, the question left unanswered is whether these new assessments approach what might be considered a truer form of reading and whether their measures of different readers performance offer representations of performance befitting the diversity and the situatedness of literacy that they tout. In other words, replacing old tests with new tests may give the impression that change is afoot, but it may simply be putting old wine into new bottles.

Some would argue that to address the problems with testing, a different epistemological approach may be needed (e.g. Moss, 1996). For example, befitting constructivist and critical research tenets, assessment would be done in a fashion that is participatory, collaborative, formative, and learner-driven. To these ends, the 1990s witnessed the rise of an orientation to assessment that involved engaging student and teachers in forms of authentic assessment, including the use of portfolios. This orientation had certain key tenets:

- Assessment should support innovative teaching and the engagement of students in strategic learning, using a range of texts for different purposes.
- Assessments should be from the inside out following from what students and teachers do rather than what the test imposes (i.e., from the outside in. Assessments should instead keep up with teaching and learning, not derail it).
- Integral to any assessment should be learning to assess oneself.

Essentially, collaborative and participatory forms of assessment such as portfolios are premised upon assessment practices being intertwined with teacher and student learning and decision-making within the classrooms or across classrooms.

In terms of benefits, these learner-based forms of assessment assumed that educators and learners would be better informed when assessments consider multiple sources of data and are done so in a collaborative, judicial, responsive, and reflexive fashion. They assumed that teachers and students have a collective understanding of achievements and progress, without the need for rigid forms of scoring and aggregation (i.e., measures derived from periodical assessments removed from the everyday). They assumed that notions of reliability can be strengthened by the verifiable nature of the evidence and grounded assumptions that could be used to generate claims. They assumed that classroom and learner-based assessments such as portfolios fit within a cultural ecological orientation that builds upon, recognizes, and values local and diverse resources and knowledges. Accordingly, they also assumed that they might bridge with, respect and credit the diverse cultural capital of communities.

Perhaps we will see a reckoning of the current resurgent interest in shifting to a socio-cultural frame with a historic political conceit. Perhaps we will also see the emergence of assessments less tied to trafficking in efficiencies that ignore diversity of our literacies and their legitimacy in the interest of supporting all (Side Comment III.6b.5).

Side Comment III.6b.5.

Our engagements in assessment have been three-fold. First, as literacy educators with an interest in how we comprehend or make meanings, We have been engaged in attempts to enlist state-of-the-art assessment techniques to study literacy—hoping to unlock the nature of meaning making and its development. Our goal has been to enlist a range of assessment tools to study reading and writing engagements precisely, reliably, and ethically. Second, we have been keen to examine the role of testing in schools and society—especially with regard to how testing practices might be enlisted by various stakeholders to make decisions. We have examined the influence and rhetoric around the growth of test-driven curriculum, and we have been involved in studies of popular tests (e.g., intelligence tests; screening tests for learning disabilities; various placement tests of reading levels; and testing procedures enlisted in judicial systems, including the interrogation of language minorities) as well as in studies of tools for reading comprehension assessment and the characteristics of large scale assessments. Third, we have been involved in helping design national and international assessments of literacy such as PISA and NAEP. Most recently, David has chaired NAEP-Reading and been guiding the plans for NAEP 2025.

As such, we have been keen to study assessment's role in terms of spurring positive, generative, organic, and culturally-responsive changes, via forms of evaluation, that empower teachers, students and caregivers—perhaps even interrupt and challenge testing's historical roots. Moving forward, we would like us to purse a course which is more culturally-oriented; to step outside of the box and advance alternatives; and to engage in research and development on more credible forms of assessment for all.

References

Afflerbach, P. (2005) High stakes testing and reading assessment. *Journal of Literacy Research*, 37 (2), 151-162.

Barr, M., Ellis, H., Hester, H. & Thomas, A. (1988) *The primary language record*, Portsmouth, NH: Heinemann.

Bruner, J. S. (1990). *Acts of meaning.* Cambridge, MA: Harvard University Press.

Calfee, R. & Hiebert, E. (1991). Classroom assessment in reading In R. Barr, M. L. Kamil, P. Mosenthal, and P. D. Pearson (Eds.). *Handbook of reading research, Volume II* (pp. 280-309). New York: Longman.

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning, 2*(1), 67–90. doi: 10.1016/0149-7189(79)90048-X.

Clay, M. (1993) *An observation survey of early literacy achievement*. Portsmouth, NH: Heinemann.

Cunningham, J., & Tierney, R. J. (1979). Evaluating cloze as a measure of cognitive change due to reading. *Journal of Reading Behavior*, *11*, 287–292.

Ellwein, M. C., Glass, G. V., Smith, M. L. (1988) Standards of competence: propositions on the nature of testing reforms. *Educational Researcher*, 17,8, 4-9.

Farr, R. & Carey, R. (1986) *Reading: what can be measured* (2nd ed.). Newark, DE: International Reading Association.

Fuchs, D., & Fuchs, L. S. (2006). Introduction to response to intervention: What, why, and how valid is it? *Reading Research Quarterly*, *41*(1), 93–99. doi:10.1598/RRQ.41.1.4.

Good, R. H., III, Kaminski, R. A., Simmons, D., & Kame'enui, E. J. (2001). Using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in an outcomes-driven model: Steps to Reading Outcomes. *Oregon School Study Council (OSSC) Bulletin*, *44*(1), 1–24. Retrieved from https://files.eric.ed.gov/fulltext/ED453526.pdf.

Goodman, K. S. (Ed.) (2006). *The truth about DIBELS: What it is, what it does*. Portsmouth, NH: Heinemann.

Gould, S. J. (1981) *The mismeasure of man*. New York: W. W. Norton.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage.

Gutiérrez, K. (2004). Literacy as laminated activity: Rethinking literacy for English learners. In J. Worthy, B. Maloch, J. V. Hoffman, D. L. Schallert, & C. M. Fairbanks (Eds.), *53rd Yearbook of the National Reading Conference* (pp. 101–114). Oak Creek, WI: National Reading Conference.

Haney, W. (1993). Testing and minorities. In L Weis And M. Fine (Eds.) *Beyond silenced voices* (pp.45-74). Albany: SUNY.

Irwin, P. A. & Mitchell, J. N. (1983). A procedure for assessing the richness of retellings. *Journal of Reading, 26*, 391-396.

Johnston, P. (1984). Assessment in reading. In P. D. Pearson, R. Barr, M. Kamil & P. Mosenthal (Ed.), *Handbook of research in Reading* (pp. 147- 182). New York: Longman.

Johnston, P. (1997). *Knowing literacy: Constructive Literacy Assessment*. York, ME: Stenhouse Publishers.

Lather, P. (1986). Research as praxis. *Harvard Educational Review, 56*(3), 257–278. doi: 10.17763/haer.56.3.bj2h231877069482.

Lemann, Nicholas (1999) The big test: the secret history of American meritocracy. New York: Farrar, Straus & Giroux

Morrow, L. M. (1988) Retelling stories as a diagnostic tool. In S. M. Glazwer, J. W. Searfoss and L. M. Gentile (Eds.), *Reexamining reading diagnosis: New trends and practices* (pp. 128-149). Newark, DE: International Reading Association.

Moss, P. (1996). Enlarging the dialogue in educational measurement: voices from interpretive research traditions. *Educational Researcher, 25*(1), 20-28.

National Assessment Governing Board (2020). Reading Framework for the 2025 National Assessment of Educational Progress. US Department of Education. Accessed July 20, 2020 *https://www.naepframeworkupdate.org*.

Nichols, S.N. & Berliner, D.C. (2007). *Collateral Damage: The effects of high-stakes testing on America's schools*. Cambridge, MA: Harvard Education Press.

U.S. Congress Office of Technology Assessment. (1992). *Testing in American Schools: Asking the right questions*. Congress of the United States, Office of Technology Assessment.

O'Reilly, T., Sabatini, J., & Wang, Z. (2018). Using scenario-based assessments to measure deep learning. In K. Millis, D. Long, J. Magliano, & K. Weimer (Eds.), *Deep comprehension: Multi-disciplinary approaches to understanding, enhancing, and measuring comprehension* (pp. 197–208). New York: Routledge.

O'Reilly, T., Weeks, J., Sabatini, J., Halderman, L., & Steinberg, J. (2014). Designing reading comprehension assessments for reading interventions: How a theoretically motivated assessment can serve as an outcome measure. *Educational Psychology Review*, *26*, 403–424. doi: 10.1007/s10648-014-9269-z.

Pearson, P. D, Palincsar, A. S., Biancarosa, G. (2020). *Reaping the rewards: Reading for Understanding Initiative.* Washington, DC; National Academy of Education.https://naeducation.org/wp-content/uploads/2020/07/NAEd-Reaping-the-Rewards-of-the-Reading-for-Understanding-Initiative.pdf

Peterson, J., Greenlaw, J. J., & Tierney, R. J. (1978). Assessing instructional placement with the I.R.I.: The effectiveness of comprehension questions. *Journal of Educational Research*, *5*, 244-250.

Resnick, D. P. & Resnick, L. P "Understanding Achievement and Acting to Produce It: Some Recommendations for the NAEP," in Phi Delta Kappan, Vol. 69, No. 8 (April 1988), pp. 576-79

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform in changing assessments. In B. K. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 37-75). Boston: National Commission on Testing and Public Policy.

Riedel, B. W., & Samuels, S. J. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students [with commentary]. *Reading Research Quarterly, 42*(4), 546–567. doi: 10.1598/RRQ.42.4.5.

Sabatini, J., O'Reilly, T., Weeks, J., & Steinberg, J. (2016, April). *The validity of scenario-based assessment: Empirical results.* Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Sabatini, J., O'Reilly, T., Weeks, J., & Wang, Z. (2019). Engineering a 21st century reading comprehension assessment system utilizing scenario-based assessment techniques. *International Journal of Testing*. doi: 10.1080/15305058.2018.1551224.

Sabatini, J. P., O'Reilly, T., Halderman, L. K., & Bruce, K. (2014a). Integrating scenario-based and component reading skill measures to understand the reading behavior of struggling readers. *Learning Disability Research & Practice, 29*(1), 36–43. doi: 10.1111/ldrp.12028.

Sabatini, J. P., O'Reilly, T., Halderman, L., & Bruce, K. (2014b). Broadening the scope of reading comprehension using scenario-based assessments: Preliminary findings and challenges. *L'Année psychologique, 114*, 693–723. doi:

Shanahan, T., Kamil, M., Tobin, A. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly*, *17*, 229-255.

Shavelson, R, Baxter, G. P, Pine, J. (1992) Performance assessment: political rhetoric and measurement reality. *Educational Researcher*, 21, 4, 22-27.

Shepard, L. & Bliem C.L. (1995). Parents' thinking about standardized tests and performance assessment. *Educational Researcher*, *24*(8), 25-32.

Shepard, L., Hannaway, J., & Baker, E. (Eds.) (2009). *Standards, Assessments, and Accountability* [Education Policy White Paper]. National Academy of Education, Working Group on Standards, Assessments, and Accountability. Retrieved from https://files.eric.ed.gov/fulltext/ED531138.pdf.

Stowell, L. P., & Tierney, R. J. (1994). Portfolios in the classroom: What happens when teachers and students negotiate assessment? In R. Allington & S. Walmsley (Eds.), *No quick fix: Rethinking literacy lessons in America's elementary schools* (pp.78-94). New York, NY: Teachers College Press.

Taylor, W. L. (1953). Cloze procedure; a new tool for measuring readability. *Journalism Quarterly*, *30*, 415-453.

Teasdale, R., Tierney, R. J., Ames, W., & Wray, R. (1978). A cross-cultural comparison of item analysis data on the Revised ITPA. *Australian Psychologist*, *3*, 391-399. https://www.tandfonline.com/doi/abs/10.1080/00050067808254328?journalCode=taps20

Tierney, R. J. (1998). Literacy assessment reform: Shifting beliefs, principled possibilities, and emerging practices. *The Reading Teacher*, *51*(5), 374–390.

Tierney, R. J. (2009). Literacy education 2.0: Looking through the rear view mirror as we move ahead. In J. Hoffman & Y. Goodman (Eds.), *Changing literacies for changing times: An historical perspective on the future of reading research, public policy, and classroom practices* (pp. 282–300). New York: Routledge.

Tierney, R. J., Carter, M., Desai, L. (1991). *Portfolio assessment in the reading writing classroom.* Norwood, MA: Christopher Gordon Publishers, Inc. https://independent.academia.edu/RobTierney/Books

Tierney, R. J., Clark, C., Fenner, L., Herter, R. J., Simpson, C. S., & Wiser, B. (1998). Portfolios: Assumptions, tensions, and possibilities. *Reading Research Quarterly*, *33*(4), 474–486. doi: 10.1598/RRQ.33.4.6.

Tierney, R. J., Crumpler, T. Bond, E., and Bertelsen, C. (2003). *Interactive assessment: Teachers, students and parents as partners.* Norwood, MA: Christopher Gordon Publishers, Inc. https://independent.academia.edu/RobTierney/Books

Tierney, R. J., & Thome, C. (2006). Is DIBELS leading us down the wrong path? In K. S. Goodman (Ed.), *The truth about DIBELS: What it is, what it does* (pp. 50–59). Portsmouth, NH: Heinemann.

Tierney, R. J., Wile, J. M., Moss, A. G., Reed, E. W., Ribar, J. P., & Zilversmit, A. (1993). *Portfolio evaluation as history: A Report on the evaluation of the history academy for Ohio teachers* [Occasional Paper]. National Council for History Education, Inc. Retrieved from https://files.eric.ed.gov/fulltext/ED371978.pdf.

Valencia S. W. & Pearson, P. D. (1987). Reading assessment: A time for change. *The Reading Teacher*, *40*, 726-732

Valencia, S. W. & Wixson, K. (2000). Policy-oriented research on literacy standards and assessment In M. L. Kamil, P. Mosenthal, R. Barr and P. D. Pearson (Eds.). *Handbook of reading research, Volume III* (pp. 909-935). New York: Longman.

Willis, A. (2008). *Reading comprehension research and testing in the US: Undercurrents of race, class, and power in the struggle for meaning.* Mahwah, NJ: Lawrence Erlbaum.