MARCIA A. INVERNIZZI
TIMOTHY J. LANDRUM
JENNIFER L. HOWELL
HEATHER P. WARLEY

Toward the peaceful coexistence of test developers, policymakers, and teachers in an era of accountability

Assessments can be technically sound in ways that preserve the theoretical integrity of reading development and provide the flexibility and instructional transparence that teachers need.

In the past five years U.S. teachers have witnessed unprecedented political insistence on the use of research-based, scientifically proven assessments and instructional techniques. Pressure to apply scientific methodology to the day-to-day work of teaching children to read and write is perhaps even stronger today than it was during the days of Sputnik and the Cold War. Then, the press for science was propelled by competition against the Soviet Union to maintain the United States's academic edge in the rush to the moon and beyond. Today, the press for science is from within, driven by the desire to preserve and extend the effectiveness of public education to all segments of our democratic society—so that "no child will be left behind." This is indeed a valiant goal, and to reach it the federal government has made money available through grants such as Reading First, designed to support states' efforts to implement scientifically based reading instruction driven by valid and reliable assessments. But there may be a disconnect between what is required to meet external demands for scientifically based reading assessment and the type of assessment information teachers need on a day-to-day basis to provide appropriately designed and targeted reading instruction for all students. In other words, there may be significant challenges associated with selecting assessment tools and implementing a comprehensive yet efficient assessment program that (a) meets high standards of scientific rigor and (b) provides teachers with instructionally useful information.

In this article we focus on the potential disconnect between research and practice in reading assessment and instruction that may be an unfortunate byproduct of increased accountability and growing emphasis on scientifically based reading research. It is important to clarify that a focus on the empirical base in designing literacy assessment and instruction is long overdue and is clearly an essential foundational step toward improved literacy—to ignore the empirical base is little short of nonsensical. Moreover, a focus on science finds generally widespread support among those concerned with enhancing the literacy development of U.S. schoolchildren. But an unintended side effect of a headlong rush toward science and accountability in assessment that does not take into account the practicalities of everyday teaching may create a disconnect between what assessments tell us about students' performance and what teachers need to know to instruct them.

We discuss specifically eight standards for the evaluation of educational assessments and assessment practices, recommended by the American Educational Research Association, the American

 doi:10.1598/RT.58.7.2

Psychological Association, and the National Council on Measurement in Education, as applied to reading assessments typically used in Reading First schools across the country. These eight standards constitute the basic obligations of test developers and research professionals to provide technically sound assessment tools to teachers. For each standard, we briefly define the constructs it entails, provide examples, and explain its importance. Next, we describe the tension that may develop in implementing each construct within classrooms, especially in the context of existing curricula and teachers' current knowledge, skill sets, and beliefs about reading. Finally, we offer solutions to these tensions by describing some examples of instructionally transparent assessments that also meet scientific requirements for technical adequacy.

Standard 1: Validity

A well-constructed assessment must first and foremost be valid. In simple terms, a measure is valid to the extent to which it measures what it is intended to measure. There are several forms of validity: content validity, construct validity, and criterion-related validity (predictive and concurrent). Relative to this discussion, predictive validity is probably the most critical focus for test developers because they must ensure that their assessments accurately predict real reading outcomes. To establish the predictive validity of an instrument, test developers compare student performance on their assessment with some external measure obtained at a later point in time. While predictive validity is critical for test developers and policymakers, teachers want assessments that are instructionally useful in the here and now.

The disconnect between scientifically based standards for assessment and the information teachers need for instruction becomes apparent in examining specific assessments that do have a high degree of predictive validity. Consider the example of tasks involving nonsense word reading. Measures of nonsense word reading are highly predictive of overall reading achievement at future points in time (Good, Wallin, Simmons, Kame'enui, & Kaminksi, 2002; Speece, Mills, Ritchey, & Hillman, 2003), and these measures appear in several reading assessments commonly used in Reading First projects across the country such as the Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1999) and Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002). The developers of such tests meet rigorous standards for predictive validity by including a nonsense word task on an assessment, and from an accountability or research standpoint, such tasks provide quite valuable information. Indeed, teachers do want to know which students are at risk for reading difficulties and in need of extra help. But a student's performance on a nonsense word task is not instructionally transparent to most teachers. For the purposes of planning and guiding instruction, teachers need specific information about a student's performance at that moment in time, such as which phonics features a first-grade student already knows and which features he or she needs to know next.

The best assessments provide tasks with a high degree of predictive validity while simultaneously providing instructionally relevant information. An example of such a task is a simple spelling assessment organized by phonics features or orthographic patterns. Spelling-by-stage inventories tell teachers where their students are along a developmental continuum of phonics and spelling achievement and exactly which phonics and spelling features a student has mastered and not mastered (Bear, Invernizzi, Templeton, & Johnston, 2004; Ganske, 2000; Viise, 1994). Grade-level spelling inventories yield instructional levels of spelling achievement and also indicate which phonics and spelling features are not yet fully developed (Henderson, 1990; Schlagal, 1986). Quantitative scores from qualitative spelling and phonics inventories have been shown to be excellent predictors of future reading achievement (Ehri, 2000; Ellis & Cataldo, 1992; Morris & Perney, 1984; Zutell & Rasinski, 1989), and they also provide information to the teacher about what phonics and spelling elements the student has already learned and which elements should be taught next. Teachers need this information for individual students as well as for the entire class to group for appropriate phonics and spelling instruction.

Standard 2: Reliability

Reliability refers to the consistency with which a test measures a construct, or the extent to which

an obtained score can be trusted to represent a "true" score. Just as there are many forms of validity, there are also many forms of reliability: test-retest, equivalent forms, split-half, and interrater. Interrater reliability is especially critical when item scoring involves a subjective judgment, such as rating the fluency of a child's oral reading on a scale of one to four, as does the National Assessment of Educational Progress (NAEP; Pinnell et al., 1995).

Reliability is important because it ensures that teachers receive accurate, trustworthy information. Because of the importance of having reliable measures, test developers sometimes may limit or avoid the use of more authentic, qualitative, or subjective measures, the reliability of which is difficult to establish, in favor of more contrived, quantitative, objective measures that can be more easily constructed to be reliable.

The measurement of reading comprehension, for example, is exceedingly complex and the currently available measures have been criticized on a number of grounds, including their inability to represent the abstract nature of the comprehension process, lack of standardized assessment strategies, and inadequate evidence of reliability and validity (Rathvon, 2004). Adequate assessment of comprehension would require multiple measures to address all of the variables in play: attention and engagement, interest, readability, vocabulary, background knowledge, oral language comprehension, written word recognition, and knowledge of genre, to name only a few (Sweet & Snow, 2003). Because administering a reliable battery of this many measures is not generally feasible for a classroom teacher or even a reading specialist, test developers usually opt for multiple choice formats that are quick and easy and that can be constructed reliably, even if narrowly. In this case, content validity (an indicator of the extent to which the questions actually measure reading comprehension) may be sacrificed for internal consistency (a measure of the reliability of the items).

Comprehension is the ultimate goal of reading, so it is a skill that teachers want to assess accurately and quickly. As a result, teachers have become dependent on the practice of asking students a few open-ended questions after the student reads a passage, or asking students to provide an oral retelling of what was just read to measure comprehension. For example, the Qualitative Reading Inventory–3 (QRI–3; Leslie & Caldwell, 2001) offers story retelling as well as open-ended explicit and implicit comprehension questions following each passage. Although nonstandardized procedures for story retelling and open-ended comprehension questions have scant evidence of validity and reliability (Rathvon, 2004), the QRI–3 and the Developmental Reading Assessment (DRA; Beaver, 1997) suggest moving students back to a reading level at which they answer most of the questions or retell most of the story correctly. Despite the unreliable nature of constructed responses, these assessments use story retelling and open-ended comprehension questions to determine a student's overall reading level, potentially limiting a student's further growth by holding them back in easier text levels even if they can read accurately and fluently at higher levels. In these instances, a technically questionable measure (i.e., open-ended comprehension questions) appears to trump more reliable measures (e.g., oral reading accuracy) in the designation of overall reading levels (Beaver; Leslie & Caldwell, 2001).

While there are other issues to consider in terms of establishing instructional reading levels, in terms of reliability, it may make more sense to focus on the more technically sound procedures available. It may be surprising to many teachers that some of the most valid and reliable measures of a student's overall reading level include simple measures of word-recognition accuracy and speed, in and out of context (Rasinki, 2000; Torgesen, Wagner, Rashotte, Burgess, & Hecht, 1997). Automatic word recognition makes reading comprehension possible. As decoding and word-recognition skills become automated, the mental capacity available for comprehension increases (LaBerge & Samuels, 1974). Once a student's reading level has been determined based on reliable measures (e.g., word recognition, oral reading accuracy, and reading rate), the student can be instructed at that level with an emphasis on comprehension instruction through building vocabulary, activating prior knowledge, scaffolding the use of comprehension strategies, and discussing what has been read.

Standard 3: Test development

Test developers are obligated to state the purpose of a given assessment tool, provide the theoretical framework, and demonstrate the technical

adequacy of their instrument through a description of the test development procedures, item analyses, field tests of revisions, and the validity of their scoring procedures. These aspects of test development may be easily overlooked because they seem obvious; however, it is important for teachers to be aware of the criteria by which test developers define an instrument. Guidelines include stating the purpose(s) of the test; defining a framework for the test; developing test specifications; and describing the process involved in developing and evaluating the items and their associated scoring procedures, assembly, and revisions.

Educators are on shaky ground when a test is used in a way or for a purpose other than that for which it was intended. If a test has been developed specifically for a certain population (e.g., preschool students, native English speakers), then it is imperative that the test be used solely for those it was designed to assess. Teachers should be careful not to use assessments designed for specific groups of students on other, more diverse, groups of students. When reviewing an instrument, decision makers should look for information about the population sampled in item development and field tests, pilot testing, and the establishment of norms. Several of the assessments that appeared on the initial Reading First list of acceptable assessment instruments, for example, were designed for students with language impairments and have questionable applicability to more diverse groups of students. Conversely, some assessments have been normed on typically achieving students and few if any special education students may have been included in the sample. This may be especially troublesome when assessing students with disabilities (e.g., learning disabilities or speech or language disorders); teachers would be wise to carefully consult test manuals to determine whether a given test is appropriate for a certain subpopulation of students.

A disconnect also may be seen when assessments are used for purposes other than those for which they were designed. A screening tool is not intended to provide diagnostic information, for example, and outcomes should not be assessed using tools not intended to measure outcomes. Reading First in fact requires that states use measures that address four purposes of assessment: screening, diagnosis, progress monitoring, and outcome assessment. However, it is not necessary that separate assessment tools be used for each purpose. For example, both the Phonological Awareness Literacy Screening (PALS; Invernizzi, Meier, & Juel, 2003; Invernizzi, Swank, Juel, & Meier, 2003) and the Texas Primary Reading Inventory (TPRI; 2003-2004) can be used for screening and then for obtaining more detailed diagnostic information. Decision makers at the district and state levels must carefully map out the various purposes of assessment in the process of selecting assessment tools.

Standard 4: Fairness in testing

Test developers are also obligated to demonstrate that their test is fair and free of bias. Fairness in testing demands the equitable treatment of all test takers. This means that tests should be free of bias in content, materials, and administration procedures that might differentially affect the performance of subgroups of test takers. Proof of this lack of bias is that students of similar levels of achievement should earn similar scores regardless of group status (i.e., gender, disability, race, ethnicity, socioeconomic status). Test developers test their measures and procedures for bias through their sampling procedures and through item analyses across different demographic segments of the population. Thus, samples used for field-testing and pilot studies should include students from all segments of the population. PALS (Invernizzi et al., 2003), for example, describes how pilot samples used in item development and field-testing mirrored the demographics of state enrollment in kindergarten through third grade.

Test developers address other elements of bias by standardizing administration procedures to prevent the subjectivity of a test administrator from unfairly swaying test results. For example, a teacher who believes a child is reading on Guided Reading level G (Fountas & Pinnell, 1999) may decide to take a running record of the child's oral reading at a level corresponding only to level G, despite the fact that the child may be able to read equally well on level H, I, or J. In this case, the absence of an objective, standardized procedure for selecting which level passage to administer for the running record may simply confirm the teacher's initial bias. While teacher knowledge of student performance provides valuable information for instruction, this prior knowledge can also result in a

failure to fully explore all possibilities in an assessment context. This type of bias, called confirmation bias (Evans, 1989), has been well documented and leads to the type of measurement selection problem described above.

A fairer way to select passages for running records would be to use an objective, standardized procedure. For example, teachers using the QRI–3 (Leslie & Caldwell, 2001) administer the passage corresponding to the highest grade-level word list on which the student achieved a score of 90% or greater. In a similar manner, teachers using the TPRI (2003-2004) first administer screening word lists that direct them to the appropriate level passage. PALS (Invernizzi et al., 2003), using a similar procedure, reported that 97% of students who read 15 or more words on a grade-level word list read the corresponding grade-level passage with 90% accuracy or greater. By providing an objective procedure for administering the passages, PALS, the QRI–3, and the TPRI avoid this common type of confirmation bias.

Standard 5: Scales, norms, and score comparability

The interpretation of test scores can be a complex task, and a full description of the different types of scores and their uses is beyond the scope of this article. Perhaps what is more important is that teachers should be familiar with basic distinctions between the most common types of tests: norm referenced and criterion referenced. Put simply, a norm-referenced test uses the results from a large and representative sample of other similar students who also took the test to establish a student's relative standing compared to his or her age- or grade-level peers. Norm-referenced scores are often expressed as percentiles; for example, a student obtained a score of 13 on a word-recognition test, which puts him or her at the 70th percentile. This means that the student's score of 13 was equal to or better than that of 70% of children the same age who took this test. It is important that teachers and other decision makers understand that norms are established by test developers who draw their normative samples from the larger population. The extent to which a given normative sample is representative of the nation as a whole, a particular state, or an individual school varies. When decisions are made based on norm-referenced testing, close attention to how norms were established is essential. Some proponents of curriculum-based measurement, for example, advocate the development and use of local norms, so that students' performances are compared to other students like them from the same school district or even the same school (e.g., Marston & Magnusson, 1988).

Tests may also be criterion referenced, which means that students' performances are measured against some established criterion, rather than against other students' performances. In the previous example, the student's score of 13 on a word list (which put that student at the 70th percentile) may still be below the criterion of 15 (out of 20) for on-grade-level reading determined by the theoretical construct of instructional level.

Researchers, administrators, and policymakers often prefer norm-referenced tests because these results indicate the standing of a student within a population of students at specific age or grade levels. Teachers, however, usually prefer the more specific instructional information provided by criterion-referenced tests. The disconnect between research-based standards for assessment practices and the culture of teaching and learning is perhaps most dramatic in the administration and scoring of norm-referenced curriculum-based measures. For example, DIBELS (Good & Kaminski, 2002) directs a teacher to discontinue students from further administration of their curriculum-based, one-minute oral reading fluency task if they read fewer than 10 words correctly on the first of three grade-level passages. While the rule for discontinuing the student informs the administrator that the student is well below expectations, it tells the teacher very little beyond what he or she already knew—that the student was struggling. Student reading of frustration-level text affords little instructional information for teachers who want to determine the student's instructional level. In this situation, a criterion-referenced test that provides multiple gradations of easy oral reading passages would allow the teacher the leeway to move up or down in passage difficulty to find the highest difficulty level at which the student can read.

It is important to note that curriculum-based measurement (CBM; Deno, 1985) presents a unique case in terms of the assessment information it provides. By definition, CBM involves

regular, brief assessments of a student's achievement within the curriculum in use. Measurement items (typically passages in the case of reading) are drawn directly from instructional materials. As such, CBM can provide useful instructional information about the effects of instruction on student performance and progress over time toward a specific criterion or goal, such as grade-level reading (e.g., Deno, 1985). An additional benefit of CBM can be derived from the establishment of local norms (Deno, 2003; Marston & Magnusson, 1988), which allow comparison of an individual child's performance with similar peers in the same class, school, or district. In this application, CBM addresses the key element of norm-referenced assessment. Thus, while they do not provide a full diagnostic picture, CBM procedures can provide teachers with tools for monitoring progress toward criterion within a normative context.

Standard 6: Standardized administration, scoring, and reporting

Standardized procedures for administering and scoring a test and for reporting its results are also essential for ensuring the accuracy and integrity of the assessment outcomes. How scores are reported clearly influences their usefulness to teachers. Some assessments report scores in terms or in formats that are intended more for statisticians than for teachers (e.g., stanines). Other assessments report results in categorical terms such as "High risk—Danger!" (Torgesen, 2003). Reports that label children in such terms might be useful for administrators, but they do not provide teachers with specific information to guide their teaching. Often raw scores from reading assessments are reported in massive data files or complicated charts intended for technical specialists or researchers who can use these files to answer specific questions of interest, often for an entire school district. Classroom teachers and reading specialists usually do not have the time and resources to interpret such massive amounts of data. Fortunately, many tests provide interpretive reports via the Internet (e.g., DIBELS, Good & Kaminski, 2002; PALS, Invernizzi et al., 2003) while also providing the option of downloading raw data files. These reports are usually in a standard format and "based on a combination of empirical data and expert judgment and experience" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 62). Rigorous computer-generated reports provide an efficient way for educators to review test scores. The PALS Internet database, for example, provides a number of interpretive reports, including class or school groupings of students by common reading levels and phonics features (Partridge, Invernizzi, Meier, & Sullivan, 2003).

Standard 7: Testing individuals of diverse linguistic backgrounds

Students whose native language is not English present a particular challenge for test developers and educators alike. According to the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999), test developers should pay special attention to "issues related to language and culture when developing, administering, scoring, and interpreting test scores and making decisions based on test scores" (p. 91). This means that norms that are based on native English speakers' performances should not be used as a comparison group for nonnative speakers. In terms of reading, most researchers agree that it is important to establish which language is dominant for a particular student, and then to establish a student's proficiency level with literacy fundamentals in that language (American Educational Research Association et al.). Although a student may not be able to read or even name the letters of the alphabet in English, he or she may be quite knowledgeable about phonetics in his or her native language. Ideally, reading assessments for individuals of diverse linguistic backgrounds should be conducted in the dominant language first, and in English as soon as English proficiency allows. Several reading assessments used in Reading First projects in the United States offer assessment in both Spanish and English. For example, Texas offers reading assessment through the TPRI (2003-2004) and a Spanish language assessment through El Inventorio de Lectura en Espanol de Tejas (Tejas LEE, 2003-2004). The problems schools face in assessing the reading development of students of diverse linguistic backgrounds are (a) the huge number of different languages now represented in the United States and the lack of corresponding assessment tools in

those languages, and (b) lack of a standardized procedure for when to include English language learners in English-literacy assessments.

One possible solution is to use a continuum of English proficiency and designate an agreed-upon level above which students are assessed in English. In Virginia, for example, students are designated along a continuum from emergent to proficient with respect to oral language, reading, and writing. The student's English-language proficiency level dictates whether he or she is included in the statewide English reading assessment.

Standard 8: Responsibilities of policy decision makers

The *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999) stated that test results are used for multiple purposes: to evaluate student achievement and growth in a domain, diagnose student strengths and weaknesses, plan educational interventions, design individual instructional plans, and place students in appropriate educational programs. To accomplish these purposes, policymakers sometimes require multiple tests to be administered. Often the tasks on tests administered to the same students overlap, causing a redundancy of testing. For example, school divisions may administer the DRA (Beaver, 1997), the Developmental Spelling Assessment (DSA; Ganske, 2000), and PALS (Invernizzi et al., 2003). When students take all three assessments, they are taking multiple oral reading in context tasks and two different qualitative spelling inventories. Furthermore, a school receiving a Reading First grant may administer a screening test (e.g., Essential Skills Screener; Erford, Vitali, Haas, & Boykin, 1995), a diagnostic test (e.g., Gray Diagnostic Reading Test; Bryant, Wiederholt, & Bryant, 2004), an additional test to monitor progress (e.g., TOWRE; Torgesen et al., 1999), and yet another test to assess outcomes (e.g., Stanford–10 Achievement Test). In addition, many schools have district- or state-imposed assessments or may be tied to earlier investments in other assessment routines. Schools may have selected several different instruments to assess the five core areas that must be assessed under Reading First (i.e., phonological awareness, phonics, fluency, vocabulary, and comprehension), in addition to meeting the four areas of assessment required (i.e., screening, diagnosis, progress monitoring, and outcomes assessment).

To decrease testing time and increase instruction time, a comprehensive, flexible assessment could be selected that evaluates all of the components of reading and serves multiple purposes. Such an assessment would ideally serve as a general screening tool yet provide an opportunity for further diagnosis on a case-need basis. PALS 1–3 (Invernizzi et al., 2003), for example, begins with a brief screening battery consisting of graded word lists and a phonics and spelling inventory. Students meeting the entry-level benchmarks established for their grade level need not be assessed further. Students who do not meet the entry-level or screening benchmarks are further diagnosed with regard to more basic skills. (Graded reading passages are also available for teachers to take running records of students' oral reading accuracy, rate their oral expression, obtain an overall reading speed, and probe comprehension.) Assessment tools that allow teachers to match students to the proper level of texts for instruction and to plan appropriate phonics instruction would serve children best.

Information teachers can use

Does the increased focus on accountability and scientific rigor mean that teachers will not get the instructionally useful information they need from newly designed assessment protocols? It certainly does not have to. Teachers have been making their own formative assessments for teaching for years. Running records of students' oral reading accuracy, miscue analyses, qualitative analyses of students' uncorrected writing samples, and the like have traditionally provided the link between assessment and instruction. While these teacher-made assessments have not always had their technical adequacy established, teachers have relied on them because they make sense, are closely tied to instruction, and reflect their beliefs about reading. They trust that these classroom-based reading assessments are valid—that they measure what they are supposed to measure. But as researchers have frequently pointed out, there is a trade-off between validity and reliability because the most reliable measures are often the narrowest, and the narrowest

measures are often the least valid. The tension that exists between scientifically based research standards for assessment versus more grass-roots utilitarian practices of teachers may be summed up by this "validity dilemma" (MacGinitie, 1993, p. 558). As valid as running records appear to be, are all teachers in the school counting the same things as errors? Are they all administering comprehension questions the same way (e.g., look back versus no look back)? Given the same array of student data such as oral reading accuracy, words per minute, and comprehension scores based on eight open-ended questions, would more than one teacher come to the same conclusion about a student's instructional reading level? Would two different teachers in the same school rate and interpret the same writing sample the same way? Research suggests that the answer to these questions is often no, and, as a result, external assessments with scientifically established validity and reliability have recently been imposed.

The result is too often a loss of instructional time and more student testing than ever before. Rather than give up what they consider to be valid, instructionally useful assessment practices such as running records of students' oral reading, most teachers have continued with their own procedures while "adding on" what is externally imposed. The resulting redundancy is staggering. It is common practice for schools with Reading First grants, for example, to administer to all students a complete informal reading inventory, a qualitative writing or spelling assessment, the assessment that comes with their new core reading program, plus externally imposed assessments required for screening, diagnosis, and progress monitoring for Reading First. In some grades, students are additionally required to take the end-of-year state standards tests in reading and math.

Some middle ground may be found in assessments that provide teachers the information they can use to teach tomorrow yet that have scientifically established evidence of technical adequacy. Such assessments must be steeped in familiar classroom practices and reflect the theoretical integrity of how children learn to read. They must be instructionally transparent and logically lead to specific instructional recommendations (Justice, Invernizzi, & Meier, 2002). In addition, ideal assessments are flexible, allowing movement within the assessment procedures to accommodate individual differences in performance. Such an assessment for grades 1 through 3 might have different levels such that not *all* students are thoroughly diagnosed. Such an assessment may have an entry level for general screening and additional deeper levels for more diagnostic information for those who did not meet the screening criteria. The information yielded from all levels should provide practical and reliable information that teachers can use: specific reading levels, phonics and spelling features, specific letter-recognition needs, letter sounds, and the like. Assessments associated with core reading programs should be carefully considered with respect to their instructional value. Components selected should be interspersed in brief intervals across time as quick probes for curricular congruency. It makes sense, for example, for teachers to periodically check students' oral reading accuracy in instructional materials to make sure they are properly placed in the right level text, or to give unannounced "spell checks" to see if certain phonics features or spelling patterns are generalizing to unstudied words. The effectiveness of all of this, however, hinges on relentless communication of purpose and procedure so that the information gleaned will be trustworthy and valid. At the same time, our current concern with policy compliance and the identification of at-risk students must be tempered with a more wholesome attempt to illustrate opportunities to help children in specific areas of literacy need.

Invernizzi teaches at the University of Virginia (125 Ruffner Hall, PO Box 400266, 405 Emmet St. South, Charlottesville, VA 22904, USA). Landrum also teaches at the University of Virginia, and Howell and Warley are research scientists there.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bear, D.R., Invernizzi, M., Templeton, S., & Johnston, F. (2004). *Words their way: Word study for phonics, vocabulary, and spelling instruction* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Beaver, J. (1997). *Developmental reading assessment*. New York: Celebration Press.

Bryant, B.R., Wiederholt, J.L., & Bryant, D.P. (2004). *Gray diagnostic reading test* (2nd ed.). Austin, TX: PRO-ED.

Deno, S.L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219–232.

Deno, S.L. (2003). Developments in curriculum-based measurement. *Journal of Special Education, 37*, 184–192

Ehri, L.C. (2000). Learning to read and learning to spell: Two sides of a coin. *Topics in Language Disorders, 20*, 19–36.

Ellis, N., & Cataldo, S. (1992). Spelling is integral to learning to read. In C.M. Sterling & C. Robson (Eds.), *Psychology, spelling, and education* (pp. 122–142). Clevedon, UK: Multilingual Matters.

Erford, B., Vitali, G., Haas, R., & Boykin, R. (1995). *Essential skills screener*. East Aurora, NY: Slosson.

Evans, J. (1989). *Bias in human reasoning: Causes and consequences*. Hillsdale, NJ: Erlbaum.

Fountas, I.C., & Pinnell, G.S. (1999). *Matching books to readers*. Portsmouth, NH: Heinemann.

Ganske, K. (2000). *Word journeys: Assessment-guided phonics, spelling, and vocabulary instruction*. New York: Guilford.

Good, R.H., & Kaminski, R.A. (Eds.). (2002). *Dynamic indicators of basic early literacy skills* (6th ed.). Retrieved November, 30, 2004, from http://dibels.uoregon.edu

Good, R.H., Wallin, J.U., Simmons, D.C., Kame'enui, E.J., & Kaminski, R.A. (2002). *System-wide percentile ranks for DIBELS benchmark assessment* (Tech. Rep. No. 9). Eugene, OR: University of Oregon.

Henderson, E.H. (1990). *Teaching spelling* (2nd ed.). Boston: Houghton Mifflin.

El Inventorio de Lectura en Espanol de Tejas (Tejas LEE). (2003–2004). Austin: Texas Education Agency and the University of Texas System.

Invernizzi, M., Meier, J.D., & Juel, C. (2003). *Phonological awareness literacy screening* (PALS 1–3). Charlottesville: University of Virginia Press.

Invernizzi, M., Swank, L., Juel, C., & Meier, J.D. (2003). *Phonological awareness literacy screening* (PALS–K). Charlottesville: University of Virginia Press.

Justice, L., Invernizzi, M., & Meier, J. (2002). Designing and implementing an early literacy screening protocol: Suggestions for the speech-language pathologist. *Language, Speech, and Hearing Services in Schools, 33*, 84–101.

Laberge, D., & Samuels, S.J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*, 293–323.

Leslie, L., & Caldwell, J. (2001). *Qualitative reading inventory–3*. New York: Addison Wesley Longman.

MacGinitie, W.H. (1993). Some limits of assessment. *Journal of Reading, 36*, 556–560.

Marston, D.B., & Magnusson, D. (1988). Curriculum-based assessment: District-level implementation. In J. Graden, J. Zins, & M. Curtis (Eds.), *Alternative educational delivery systems: Enhancing instructional options for all students* (pp. 137–172). Washington, DC: National Association of School Psychologists.

Morris, D., & Perney, J. (1984). Developmental spelling as a predictor of first-grade reading achievement. *Elementary School Journal, 84*, 441–457.

Partridge, H., Invernizzi, M., Meier, J., & Sullivan, A. (2003, November/December). Linking assessment and instruction via Web-based technology: A case study of a statewide early literacy initiative. *Reading Online, 7*(3). Retrieved November 17, 2004, from http://www.readingonline.org/articles/art_index.asp?HREF=partridge/index.html

Pinnell, G.S., Pikulski, J.J., Wixson, K.K., Campbell, J.R., Gough, P.B., & Beatty, A.S. (1995). *Listening to children read aloud*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Rasinski, T.V. (2000). Speed does matter in reading. *The Reading Teacher, 54*, 146–151.

Rathvon, N. (2004). *Early reading assessment: A practitioner's handbook*. New York: Guilford.

Schlagal, R. (1986). Informal and qualitative assessment of spelling. *Pointer, 30*(2), 37–41.

Speece, D., Mills, C., Ritchey, K., & Hillman, E. (2003). Initial evidence that letter fluency tasks are valid indicators of early reading skill. *Journal of Special Education, 36*, 223–233.

Sweet, A.P., & Snow, C.E. (Eds.). (2003). *Rethinking reading comprehension*. New York: Guilford.

Texas Primary Reading Inventory. (2003-2004). Austin: Texas Education Agency and the University of Texas System.

Torgesen, J.K. (2003). *Establishing a firm foundation: Phonemic awareness and phonics*. Retrieved June 3, 2004, from http://www.justreadflorida.org/conf-03-lead/ppt-torgesen.pdf

Torgesen, J.K., Wagner, R.K., & Rashotte, C.A. (1999). *Test of word reading efficiency*. Austin, TX: PRO-ED.

Torgesen, J.K., Wagner, R.K., Rashotte, C.A., Burgess, S., & Hecht, S. (1997). Contributions of phonological awareness and rapid naming ability to growth of word-reading skills in second to fifth grade. *Scientific Studies of Reading, 12*, 161–185.

Viise, N.M. (1994). *Feature word spelling list: A diagnosis of progressing word knowledge through an assessment of spelling errors*. Unpublished doctoral dissertation, University of Virginia, Charlottesville.

Zutell, J., & Rasinski, T. (1989). Reading and spelling connections in third and fifth grade students. *Reading Psychology, 10*, 137–155.