Psychology in the Schools, Vol. 42(6), 2005
Published online in Wiley InterScience (www.interscience.wiley.com).

© 2005 Wiley Periodicals, Inc.
DOI: 10.1002/pits.20104

HOW RELIABLE ARE INFORMAL READING INVENTORIES?

JANET E. SPECTOR

University of Maine

Informal Reading Inventories (IRI) are often recommended as instructionally relevant measures of reading. However, they have also been criticized for inattention to technical quality. Examination of reliability evidence in nine recently revised IRIs revealed that fewer than half report reliability. Several appear to have sufficient reliability for lower stakes decisions such as selection of classroom reading materials, but not for higher stakes purposes such as identification of reading difficulties. This article provides recommendations for improving IRI reliability and addresses the need for expanded guidelines for evaluating reliability, particularly for measures of score agreement. © 2005 Wiley Periodicals, Inc.

Heightened awareness of the importance of timely intervention for reading difficulties has resulted in increased demand for instructionally relevant assessments of early reading skills. Instructionally relevant assessments serve not only to identify students who have not benefited fully from previous instruction, but also to direct efforts to accelerate progress. To meet this demand, school personnel have turned to a variety of program-specific and off-the-shelf measures of reading. In the latter category, authors of textbooks on reading difficulties often recommend Informal Reading Inventories (IRIs) as supplements to day-to-day classroom assessments and less frequently administered standardized tests (e.g., Barr, Blachowicz, Katz, & Kaufman, 2002; Gunning, 2002; McKenna & Stahl, 2003). Many special education assessment textbooks also identify IRIs as viable tools for evaluating reading skills in students who are referred for special education services or who receive special education services (McLoughlin & Lewis, 2005; Overton, 2003; Taylor, 2003). Indeed, a survey of special educators in four states revealed that special educators are just as likely to use IRIs as they are to use another widely recommended technique, curriculum-based measurement (CBM) of oral reading fluency (Arthaud, Vasa, & Steckelberg, 2000).

Procedurally, IRIs assess a student's instructional level in reading using sets of passages that are written or selected to be representative of the difficulty level of texts at different grade levels, and in different schools and reading programs. Most IRIs provide a choice of three to six passages per grade and examiners are free to choose which passages to use with a particular student. Assessment typically begins at a comfortable level for the student and continues until the upper limit of the student's instructional range is reached. Estimates of a student's reading level may be based on a combination of oral reading accuracy (i.e., percentage of words read correctly) and comprehension, or on comprehension alone. Comprehension is assessed primarily by post-reading questioning, although some IRIs also use retelling (i.e., free recall of the passage).

Descriptively, IRI scores are grade-referenced, indicating the highest level at which students can read with sufficient ease and accuracy to gain meaning from text. For example, based on IRI performance, a student might be identified as reading at a third-grade level for instructional purposes. Each elementary grade, except first grade, is represented by a single reading level. Because reading skills change so dramatically during the first year of formal instruction, most IRIs distinguish three levels of skill: preprimer, primer, and grade 1. Beyond the elementary years, many IRIs report reading level using more global designators such as middle level or high school.

According to Paris (2002), IRIs have been used in the past primarily for diagnostic purposes such as observation of a student's approach to decoding unfamiliar words or comparison of a

Correspondence to: Janet E. Spector, 5766 Shibles Hall, University of Maine, Orono, ME 04469–5766. E-mail: janet.spector@umit.maine.edu

student's success in reading narrative versus expository passages. At the same time, however, he reported increasing use of IRI scores for higher stakes purposes such as measuring reading growth and reporting annual progress (Paris, 2002). In the state in which the present investigation was conducted, for example, IRI results are commonly used to identify students' present level of performance in reading on Individualized Education Programs (IEPs), to evaluate year-to-year growth in reading, and to target students for Title I intervention services.

Despite their intuitive appeal, IRIs have been subject to persistent criticism for inattention to reliability. More than 20 years ago, Pikulski and Shanahan (1982) called for additional studies of alternate-forms and test–retest reliability, as well as further research to establish the reliability of criteria for determining students' instructional level for text. Klesius and Homan's (1985) review of research on IRIs also identified serious problems with respect to interrater reliability.

Unfortunately, much of the literature on uses and limitations of IRIs dates back to the 1980s and early 1990s and, consequently, it remains unclear whether previous criticisms still apply. On the one hand, researchers associated with the Center for the Improvement of Early Reading Achievement (CIERA) have studied IRIs in their work on early literacy assessment and intervention, concluding that most commercial IRIs are based on "acceptable levels" of reliability (Paris & Carpenter, 2003, p. 579). On the other hand, reviews of IRIs in the two most recent volumes of *Mental Measurement Yearbook* indicate that reliability evidence remains sparse (Burns, 2003; Gratz, 2003; Shanahan, 2001; Stahl, 2001). The discrepancy between these sources suggests the need for an updated investigation of reliability among current IRIs.

Furthermore, most previous IRI research has been disseminated in sources that cater to teachers. School psychologists and educational diagnosticians may be less familiar with the uses and limitations of IRIs. Given Paris's (2002) report of a resurgence of interest in IRIs, it is important for all professionals who are charged with early identification of reading difficulties to be current in their knowledge about these instruments. Access to up-to-date analyses is particularly critical for school psychologists who play a leadership role on school-based assessment teams and advise teachers regarding test selection.

The purpose of the present investigation was to update the knowledge base on IRI reliability and to identify critical issues in evaluating the reliability of IRI scores. Reliability was selected as a focus for two reasons. First, as mentioned above, previous IRI evaluations have consistently identified reliability as a weakness (Klesius and Homan, 1985; Pikulski & Shanahan, 1982). Second, reliability is generally regarded as a necessary, although not sufficient, condition for validity (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). In other words, if a test does not have adequate reliability, then it is inadvisable to use that measure, even if the tasks and materials appear to be well-aligned with classroom instruction.

Method

Measures

Nine recently revised IRI manuals were examined for evidence of reliability. These instruments were identified based on a search of Education Resources Information Center (ERIC), the last several volumes of *Mental Measurements Yearbook* (Plake & Impara, 2001; Plake & Impara, 2003; Impara & Plake, 1998), recent textbooks on assessment, and the Web sites of test and textbook publishers. Informal Reading Inventories that had been published or revised since 1999 were targeted to ensure an up-to-date analysis. Consistent with the review's focus on use of IRIs to identify early reading difficulties, middle level or high school IRIs were also eliminated. This resulted in a pool of nine inventories: Analytical Reading Inventory, 7th edition (ARI; Woods &

Moe, 2003); Bader Reading and Language Inventory, 4th edition (B-RLI; Bader, 2002); Basic Reading Inventory, 8th edition (BRI; Johns, 2001); Classroom Reading Inventory, 9th edition (CRI; Silvaroli & Wheelock, 2001); Ekwall/Shanker Reading Inventory, 4th edition (ESRI; Shanker & Ekwall, 2000); Burns Roe Informal Reading Inventory, 6th edition (BR-IRI; Roe, 2002); Qualitative Reading Inventory-3 (QRI-3; Leslie & Caldwell, 2001); Reading Inventory for the Classroom, 5th edition (RIC; Flynt & Cooter, 2004); and Steiglitz Informal Reading Inventory, 3rd edition (SIRI; Steiglitz, 2002). Interestingly, a previous IRI study identified eight of the nine instruments as the most widely disseminated IRIs based on market data and citations in the literature (Applegate, Quinn, & Applegate, 2002).

Procedures

Reliability evidence in IRI manuals was located by examining the table of contents. Because few manuals identify reliability as a topic in the table of contents, evidence was also gleaned from sections that describe test rationale or test development. Information was coded using conventional categories of reliability: test–retest, alternate forms, internal consistency, and interrater.

Reliability categories. Studies of test–retest reliability examine evidence of consistency in performance on the same test passage on two different occasions, whereas studies of alternate-forms reliability investigate consistency in performance across two different forms of the same test. Analyses of internal consistency require only a single administration of a test (i.e., one occasion, one form) and reflect consistency in performance across items within the test. Studies of interrater reliability indicate the consistency among scorers in evaluating student performance.

Score consistency versus agreement. Two approaches to estimating reliability were distinguished: score consistency and score agreement. Score consistency reflects the stability of student rankings across occasions, forms, or scorers. With regard to forms, for example, do students who rank high on Form A also rank high on Form B? Conversely, do students who rank low on Form A also rank low on Form B? To achieve high score consistency, students do not need to receive the same score on each form; rather they need to maintain the same rank relative to their peers on both forms. Score consistency is typically estimated through correlational or ANOVA-based generalizability analyses (Crocker & Algina, 1986).

The score-agreement approach considers the extent of agreement across occasions, forms, or scorers in assigning students the same score. For example, if a student is identified as a second-grade reader on Form A, is this student also identified as a second-grade reader on Form B? Score agreement is estimated by computing the proportion of students who are assigned the same score across occasions, forms, or raters (i.e., point-to-point agreement) or by using more complex formulas that adjust for factors such as chance agreement (e.g., Cohen's κ) or proximity of scores to decision cut-offs (Livingston's K^2).

Results

Available evidence of reliability is summarized in Table 1. Surprisingly, only four IRI manuals provide estimates of reliability.[1] These four manuals all report alternate-forms reliability; one also reports internal consistency and interrater reliability. Issues pertaining to the quality and scope of IRI reliability studies are examined below.

[1] An additional IRI, the BRI (Johns, 2001), does not report reliability, but does reference Helgren-Lempesis and Mangrum's (1986) study of the 1981 edition of the test.

Table 1
Reliability of Estimates of Students' Instructional Level in Reading as Reported in the Manuals of Four Informal Reading Inventories

Test	Reliability type	Statistical approach	Measure	Reliability
Bader Reading	Alternate forms	Score consistency	Oral reading	
& Language Inventory		Pearson's r	Elementary	.80
			Secondary/Adults	.83
			Silent reading	
			Elementary	.78
			Secondary/Adults	.79
Ekwall/Shanker	Alternate forms	Score consistency	Oral reading	.82
Reading Inventory		Pearson's r	Silent reading	.79
Qualitative Reading	Alternate forms	Score agreement	Silent reading	>.80[a]
Inventory-3		Livingston's K^2		
Qualitative Reading	Alternate forms	Point-to-point agreement	Silent reading	
Inventory-3			Primer	71%
			Gr. 1	86%
			Gr. 2	78%
			Gr. 3	80%
			Gr. 4	80%
			Gr. 5	75%
			Gr. 6	77%
			Upper middle school	81%
	Internal consistency	Score consistency	Silent reading	___[b]
		Variance components		
Qualitative Reading	Interrater	Score consistency	Oral reading	
Inventory-3		Cronbach's α	Total miscues	.99
			Meaning-changing miscues	.99
			Silent reading	
			Explicit questions	.98
			Implicit questions	.98
Steiglitz Informal	Alternate forms	Score agreement	Oral reading	
Reading Inventory		Point-to-point agreement	Narrative passages	80%
			Expository passages	86%
			Narrative with expository passages	69%

[a] Individual coefficients are not reported; however, the test's manual indicates that all coefficients were above .80 and 75% were above .90. [b] Reliability coefficients are not reported. Instead, *SEM*s are listed for each passage.

Alternate-Forms Reliability

Alternate-forms reliability is an important consideration for IRIs because passage selection is not prescribed. That is, examiners can choose to administer any of the passages at a particular reading level. If reading level varies widely depending on choice of passage, then examiners cannot have confidence in the accuracy of test results. Among the four IRIs that report alternate-forms reliability, two provide evidence of score consistency and two provide evidence of score agreement.

Score consistency. Two test manuals, the ESRI (Shanker & Ekwall, 2000) and B-RLI (Bader, 2002) report alternate-forms reliability for both oral and silent reading. The ESRI results are based on a study of 40 students in grades 1–9, while the B-RLI results are based on studies of 40 elementary and 40 secondary students/adults. Neither manual indicates year in which data were collected; characteristics of participating students, teachers, and schools; or number of students by age or grade.

Further, both tests report reliability across, rather than within, grade. This approach is problematic for two reasons. First, the overall correlation across grades may mask significant between-grade differences in test reliability. Second, because reading scores increase over grade, reliability estimates for all grades combined are apt to be inflated (Hammill, Brown, & Bryant, 1992; Salvia & Ysseldyke, 2004). As assessment experts routinely advise, reliability of developmentally sensitive measures should be reported by age or grade (Linn & Gronlund, 2000; Salvia & Ysseldyke, 2004).[2]

These limitations notwithstanding, both sets of studies resulted in similar estimates of alternate-forms reliability, with coefficients ranging from .79 to .83 (see Table 1). Descriptively, coefficients for oral reading were marginally higher than for silent reading, a pattern that may reflect the efficacy of combining measures of word recognition and comprehension when estimating a student's instructional level.

Score agreement. Informal Reading Inventories result in absolute judgments about a student's reading level rather than judgments about a student's ranking relative to others who took the same test. For this reason, score agreement is an important consideration in evaluating the reliability of an IRI. That is, how often are students assigned the same score on alternate forms?

The SIRI's author (Steiglitz, 2002) examined point-to-point agreement in pairs of narrative (Forms A and C; $n = 20$; grades 3–11) and expository passages (Forms B and D; $n = 28$ students; grades 1–12). The manual does not indicate when data were collected or discuss characteristics of students who were tested. In each study, estimates of students' instructional level were based on oral reading and comprehension questioning. As Table 1 shows, roughly four of every five students were assigned the same score on alternate forms. Comparable results were obtained for narrative and expository forms alike.

A second SIRI study investigated the equivalence of narrative and expository passages at the same reading level. Not surprisingly, passages that differed in genre resulted in a somewhat lower rate of score agreement than passages of the same genre. As seen in Table 1, about two-thirds of the 73 students in grades 1–9 who were tested on Form A (narrative) and Form B (expository) were placed at the same instructional level on both forms. This finding suggests that control for genre has a positive effect on reliability.

The QRI-3's authors (Leslie & Caldwell, 2001) investigated score agreement using two approaches: point-to-point agreement (like the SIRI) and Livingston's K^2 (Livingston, 1972). The manual does not identify the number of students who were included in these analyses, their grade levels, or when data were collected. However, the QRI-3 pilot study included 267 students in grades 1–4 and grade 8. Instructional level scores for both analyses were based on silent reading (i.e., comprehension questioning only). Analyses compared performance on pairs of passages that had been identified by the authors as similar in genre and topic familiarity. Score agreement ranged from 71–86% (median = 79%) across eight reading levels. There appeared to be no relationship between grade and proportion of agreement (see Table 1).

[2] Interestingly, based on within-grade analyses of reliability, Helmgren-Lempesis and Mangrum (1986) reported a correlation of only .69 between alternate forms for an earlier edition of the ESRI.

Additional analyses were completed using Livingston's K^2, a procedure for evaluating agreement that takes into consideration how close scores are to the decision-making cut-off. To do so, score discrepancies between the two forms that are close to the cut-off are given less weight than score discrepancies far from the cut-off. Coefficients are not reported, although the manual states that all were above .80, with three-quarters of these coefficients being at least .90. Interpretive issues related to measures of score agreement are discussed in a later section.

Internal Consistency

The QRI-3's authors (Leslie & Caldwell, 2001) examined internal-consistency reliability in sets of post-reading comprehension questions that accompany IRI passages. Consistent with the Crocker and Algina (1986) recommendation for reporting reliability on criterion-referenced tests, they report the standard error of measurement (*SEM*) for each passage rather than reliability coefficients. The *SEM* is used to determine whether obtained comprehension scores (i.e., percentage of questions answered correctly) yield a precise estimate of the student's "true" comprehension score.

Overall, the authors note that *SEM*s for comprehension scores at all reading levels were relatively large, a finding that indicates low reliability. They illustrate the implications of this finding using the example of a student who responds correctly to 75% of the comprehension questions for the passage *Sequoyah*, which has a *SEM* of .18 (or 18%). With this *SEM*, the student's estimated true score is anywhere from 39% to $>$ 100% correct (i.e., 75% ± 2 *SEM*s). This is problematic because the QRI-3's cut-off for the instructional level is 70% comprehension. Although the student's obtained score exceeds the cut-off, the true score could fall below the cut-off. To address this problem, the QRI-3 manual recommends that teachers base estimates of the instructional level on comprehension questions from at least two passages at the same level. Unfortunately, this caveat is included in the chapter that details test development—not in the chapter that outlines administration and scoring procedures, where it more likely would be read. Further implications of this problem are discussed in a later section.

Interrater Reliability

Among nine current IRIs, only the QRI-3 manual reports interrater reliability. Three expert scorers and an unidentified number of other examiners with less extensive training evaluated 122 readings, 49 oral and 73 silent. As shown in Table 1, reliability analyses (Cronbach's α) indicated high scoring consistency for two types of oral reading errors (i.e., meaning changing and nonmeaning changing) and two types of comprehension questions (i.e., textually explicit and textually implicit). These results suggest that it is possible to train examiners to score IRIs in a way that results in similar rank-ordering of students from one tester to the next.

Discussion

More than 20 years ago, two reviews identified need for further research to establish the reliability of commercial IRIs, particularly for use in determining a student's instructional level in reading (Klesius & Homan, 1985; Pikulski & Shanahan, 1982). The present investigation was designed to determine whether subsequent editions of IRIs have addressed previous concerns. Specifically, this review examined reliability evidence in test manuals of nine current IRIs. Its purpose was to inform those who use IRIs or who oversee school-based assessment teams regarding (a) the reliability of particular IRIs, and (b) issues involved in evaluating the reliability of IRIs.

Key Findings

The present investigation yielded three major findings. First, many earlier criticisms of IRIs are as apt today as they were two decades ago. Some recent research has resulted in optimistic reports about the reliability of commercial IRIs (Paris & Carpenter, 2003); however, fewer than half of the IRIs manuals that were examined provide evidence of reliability. Given that most have been on the market for 20 to 30 years, failure to address reliability appears to reflect a considered decision by some IRI authors to ignore widely accepted professional standards of test quality (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

Poor documentation. Even among IRIs that have investigated reliability, much of the evidence is poorly documented. The majority of test authors fail to provide sufficient descriptions of research methodology or statistical results. For example, manuals do not always identify how many students were tested or at what grade levels. It is difficult to determine whether results are based on the current edition of the test because most manuals do not indicate when studies were conducted. Although earlier findings might still apply, it should not be a foregone conclusion. Informal Reading Inventories' authors are obligated to make their case for the relevance of older data and to be clear in sorting out results by test edition.

Weak research methodology. The contribution of some efforts to establish reliability is undermined by weak research methodology. Many of the studies described in IRI manuals are small scale with respect to the number of students, grades, and schools. Consequently, the generalizability of results is questionable. For example, one reliability study involved only 20 students in grades 3–11. Although such a study contributes to the knowledge base about test reliability, it is less than compelling as the principal source of evidence.

Further, some tests have established reliability across—rather than within—grade to compensate for small sample size. That is, scores from students in a broad grade span have been combined when computing correlations between alternate forms. However, because instructional reading levels increase across grades, correlations that are based on combined grades tend to be inflated. Thus in the present sample, the ESRI and B-RLI alternate-forms reliability coefficients are likely to be over-estimates.

Finally, most IRI authors have investigated only a single type of reliability: alternate forms. Although this type of reliability is critical to IRI use, interrater reliability also is important. As noted in the *Standards for Educational and Psychological Measurement*, evidence of interrater consistency should be provided whenever test scoring requires subjective judgment (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; p. 33). Only one IRI provides any evidence of interrater reliability despite the fact that previous research has identified problematic inconsistencies across examiners. This is a major shortcoming that needs to be addressed by IRI developers.

Implications for Practice

Given the above limitations, what should assessment leaders conclude about appropriate use of IRIs? Three practical implications of the present study are discussed below.

Potential for harm. First, IRIs that provide no evidence of reliability should not be used to estimate a student's reading level, regardless of how casually the results will be applied. School personnel may regard IRIs as benign tools because they are not standardized tests. Some teachers, for example, use IRIs to generate diagnostic profiles (e.g., performance on word recognition vs. comprehension of on narrative vs. expository passages). However, any test—no matter how informal—has the potential for harm if the information it provides is imprecise or misleading. In the absence of reliability studies, it is risky to assume that particular IRIs measure anything stable or generalizable about a student's reading performance. In the present pool, five IRI manuals do not report any reliability data: ARI (Woods & Moe, 2003), BR-IRI (Roe, 2002), BRI (Johns, 2001), CRI (Silvaroli & Wheelock, 2001), and RIC (Flynt & Cooter, 2004).

Consideration of intended test use. Second, for IRIs that do report reliability, examiners need to evaluate the strength of the evidence in light of the intended use of results. Although measurement error is expected on any test, the amount of error that is tolerable depends on how the test scores will be used. As noted in the most recent edition of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999):

> Precision and consistency of measurement are always desirable. However, the need for precision increases as the consequences of decisions and interpretations grow in importance. If a decision can and will be corroborated by information from other sources or if an erroneous initial decision can be quickly corrected, scores with modest reliability may suffice. But if a test score leads to a decision that is not easily reversed, such as rejection or admission of a candidate to a professional school or the decision by a jury that a serious injury was sustained, the need for or a high degree of precision is much greater. (p. 30)

How high should reliability be for IRIs? Relatively few sources provide explicit guidelines in this regard. Salvia and Ysseldyke (2004) and Nitko (2001) recommend a minimum reliability of .90 for tests that will inform high-stakes decisions about individual students (e.g., tracking, placement in special programs, graduation, admission to higher education institutions). On the other hand, a more relaxed minimum reliability of .80 is acceptable where tests are used to determine need for additional testing (Salvia & Ysseldyke, 2004). Nitko (2001) provides further guidance regarding classroom-based assessments, recommending a minimum reliability of .70 for measures that will be used for routine instructional purposes.

Two of the most common uses of IRI instructional level scores are placing students in reading materials and identifying reading difficulties (McCabe, Margolis, & Barrenbaum, 2001). The first purpose, placing students in instructional materials, most closely matches Nitko's description of routine classroom assessment. In such a situation, incorrect decisions can easily be detected and reversed with little risk of harm to students; a minimum reliability of .70 therefore seems reasonable. Available evidence, shown in Table 1, indicates that several IRIs in the present pool meet this standard, although methodological and reporting limitations discussed above necessitate a tentative, rather than firm, conclusion.

On the other hand, determining whether a student requires intervention for reading difficulties is a higher stakes decision. When a child who needs intervention is *not* identified or a child who does not need intervention *is* (incorrectly) identified, the risk of harm is considerable. Furthermore, an error in judgment might not be immediately detected or easily corrected. For this reason, a minimum reliability of .90 is desirable, even when IRI results will be used in combination with other measures. Although single tests should never be used as the sole source of evidence in educational decision-making, each test that contributes to higher stakes decisions should meet the appropriate reliability standard (Nitko, 2001).

With the exception of the QRI-3, none of the IRIs in the present pool approach this level of reliability. However, even the QRI-3 is questionable if standard administration procedures are followed. As mentioned above, the authors of the QRI-3 recommend that teachers base judgments about a student's reading level on at least two passages at the same grade level, due to poor internal-consistency reliability of comprehension scores. Unfortunately, this recommendation is not reflected in directions for administering the inventory and, therefore, many examiners will not take the necessary steps to maximize the reliability of the inventory.

Although only the QRI-3 investigated the reliability of comprehension scores, it is likely that many other IRIs share a similar problem. Indeed, several test authors note the desirability of examining performance across several passages at the same level; however, they stop short of prescribing this practice. It is likely that IRI reliability would be enhanced if examiners based decisions regarding the instructional level on multiple passages, a recommendation that has been made by others as well (McKenna & Stahl, 2003; Paris & Carpenter, 2003). To make this happen, however, IRI authors must revise their instructions for administration and scoring to make use of multiple passages the default option. In the mean time, assessment leadership personnel should advise IRI users in their schools to compute scores across at least two passages at the same grade level.

Need for expanded reliability guidelines. A third implication for practice pertains to the need for further attention within the assessment community to reliability standards. Recent federal initiatives place increased pressure on school personnel to select technically adequate tests. Unfortunately, relatively few sources provide specific guidelines to assist school personnel in evaluating evidence reported in test manuals. Guidelines provided by Salvia and Ysseldyke (2004) and Nitko (2001) are a start. However, two limitations constrain their usefulness. First, few examples are provided, making it difficult for examiners to relate guidelines to their own assessment purposes and practices. Second, the guidelines are based exclusively on score consistency—score agreement is neglected. The assessment literature does not provide guidance regarding desirable levels of score agreement. On the surface, estimates of percentage agreement may look like conventional reliability coefficients because they can be expressed as decimal values. However, they do not have the same meaning as reliability coefficients; therefore, they cannot be evaluated in the same manner (Crocker & Algina, 1986). For example, the SIRI presents results indicating agreement between alternate forms of around 80%. This is not equivalent to a conventional reliability coefficient of .80.

How high should level of agreement be across alternate forms or scorers? As was the case for score consistency, the need for high score agreement no doubt depends on the nature of decisions that will be made. For purposes of determining whether a student needs reading intervention, an error in judgment that occurs in one out of every five cases is probably too high. On the other hand, 80% agreement between forms may be sufficient if teachers plan to use the test to group students for instruction and will have the opportunity to corroborate decisions by observing subsequent performance. As schools move toward more frequent use of tests that result in absolute judgments regarding student proficiency in domains such as reading, score agreement is apt to become an increasingly critical dimension of reliability. Development of more comprehensive and explicit guidelines for evaluating test reliability is recommended to facilitate appropriate test selection by school personnel.

Conclusions

Informal Reading Inventories are intuitively appealing instruments for assessing student performance in reading. They engage students in instructionally relevant, oral reading and comprehension tasks, and they yield scores that are readily understood by classroom teachers and parents. However, IRI authors need to provide higher quality and broader-based reliability evidence to support their use. Although technical adequacy is not determined by reliability alone, without reliability, test utility is severely limited. School psychologists and other professionals who play a leadership role on school-based assessment teams must be well informed about the uses and limitations of IRIs so that they can educate teachers about these issues, guard against inappropriate use of IRIs, and assist teachers in selecting measures with adequate reliability for particular purposes.

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.

Applegate, M.K., Quinn, K.B., & Applegate, A.J. (2002). Levels of thinking required by comprehension questions in informal reading inventories. The Reading Teacher, 56, 174–180.

Arthaud, T.J., Vasa, S.F., & Steckelberg, A.L. (2000). Reading assessment and instructional practices in special education. Diagnostique, 25, 205–228.

Bader, L.A. (2002). Bader Reading and Language Inventory (4th ed.). Upper Saddle River, NJ: Merrill.

Barr, R., Blachowicz, C.Z., Katz, C., & Kaufman, B. (2002). Reading diagnosis for teachers: An instructional approach (4th ed.). Boston: Allyn & Bacon.

Berk, R.A. (1984). Selecting the index of reliability. In R.A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 231–266). Baltimore, MD: Johns Hopkins University Press.

Burns, M.K. (2003). Review of the basic reading inventory. In B.S. Plake, J.C. Impara, & R.A. Spies (Eds.), Fifteenth mental measurements yearbook (8th ed., pp. 101–103). Lincoln, NE: Buros Institute of Mental Measurements.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt Rinehart Winston.

Flynt, E.S., & Cooter, R.B., Jr. (2004). Reading inventory for the classroom (5th ed.). Upper Saddle River, NJ: Merrill.

Gratz, Z. (2003). Review of the basic reading inventory. In B.S. Plake, J.C. Impara, & R.A. Spies (Eds.), Fifteenth mental measurements yearbook (8th ed., pp. 104–105). Lincoln, NE: Buros Institute of Mental Measurements.

Gunning, T.G. (2002). Assessing and correcting reading and writing difficulties (2nd ed.). Boston: Allyn & Bacon.

Hammill, D.D., Brown, L., & Bryant, B.R. (1992). A consumer's guide to tests in print. Austin, TX: Pro-ed.

Harwell, M. (2001). Review of the basic reading inventory. In B.S. Plake & J.C. Impara (Eds.), The fourteenth mental measurements yearbook (7th ed., pp. 112–113). Lincoln, NE: Buros Institute of Mental Measurement.

Helgren-Lempesis, V.A., & Mangrum, C.T. (1986). An analysis of alternate-form reliability of three commercially-prepared informal reading inventories. Reading Research Quarterly, 21, 209–215.

Impara, J.C., & Plake, B.S. (Eds.). (1998). The fourteenth mental measurements yearbook. Lincoln, NE: Buros Institute of Mental Measurement.

Johns, J.L. (2001). Basic Reading Inventory: Pre-primer through grade twelve and early literacy assessments (8th ed.). Dubuque, IA: Kendall-Hunt Publishing Co.

Klesius, J.P., & Homan, S.P. (1985). A validity and reliability update on the informal reading inventory with suggestions for improvement. Journal of Learning Disabilities, 18, 71–75.

Leslie, L., & Caldwell, J. (2001). Qualitative Reading Inventory-3. New York: Longman.

Linn, R.L., & Gronlund, N.E. (2000). Measurement and assessment in teaching (8th ed.). Upper Saddle River, NJ: Prentice Hall.

Livingston, S.A. (1972). Criterion-referenced application of classical test theory. Journal of Educational Measurement, 9, 13–26.

McCabe, P.P., Margolis, H., & Barrenbaum, E. (2001). A comparison of Woodcock–Johnson Psycho-Educational Battery-Revised and Qualitative Reading Inventory-II instructional reading levels. Reading and Writing Quarterly, 17, 279–289.

McKenna, M.C., & Stahl, S. (2003). Assessment for reading instruction. New York: Guilford.

McLoughlin, J.A., & Lewis, R.B. (2005). Assessing students with special needs (6th ed.). Upper Saddle River, NJ: Prentice-Hall.

Nitko, A.J. (2001). Education assessment of students (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.

Overton, T. (2003). Assessing learners with special needs: An applied approach (4th ed.). Upper Saddle River, NJ: Prentice-Hall.

Paris, S.G. (2002). Measuring children's reading development using leveled texts. The Reading Teacher, 56, 168–170.

Paris, S.G., & Carpenter, R.D. (2003). FAQs about IRIs. The Reading Teacher, 56, 578–580.

Pikulski, J.J., & Shanahan, T. (1982). Informal reading inventories: A critical analysis. In J.J. Pikulski & T. Shanahan (Eds.), Approaches to informal evaluation of reading. Newark, DE: International Reading Association.

Plake, B.S., & Impara, J.C. (Eds.). (2001). The fourteenth mental measurements yearbook. Lincoln, NE: Buros Institute of Mental Measurement.

Plake, B.S., & Impara, J.C. (Eds.). (2003). The fifteenth mental measurements yearbook. Lincoln, NE: Buros Institute of Mental Measurement.

Roe, P.C. (2002). Burns Roe Informal Reading Inventory: Preprimer to twelfth grade (6th ed.). Boston: Houghton Mifflin.

Salvia, J., & Ysseldyke, J.E. (2004). Assessment (9th ed.). Boston: Houghton-Mifflin.

Shanahan, T. (2001). Review of the Burns/Roe Informal Reading Inventory). In B.S. Plake & J.C. Impara (Eds.), The fourteenth mental measurements yearbook (5th ed., pp. 196–198). Lincoln, NE: Buros Institute of Mental Measurement.

Shanker, J.L., & Ekwall, E.E. (2000). Ekwall/Shanker Reading Inventory (4th ed.). Boston: Allyn and Bacon.

Silvaroli, N.J., & Wheelock, W.H. (2001). Classroom Reading Inventory (9th ed.). Boston: McGraw-Hill Higher Education.

Stahl, S. (2001). Review of the Basic Reading Inventory. In B.S. Plake & J.C. Impara (Eds.), The fourteenth mental measurements yearbook (7th ed., pp. 113–114). Lincoln, NE: Buros Institute of Mental Measurement.

Steiglitz, E.L. (2002). The Steiglitz Informal Reading Inventory: Assessing reading behaviors from emergent to advanced levels (3rd ed.). Boston: Allyn and Bacon.

Taylor, R.L. (2003). Assessment of exceptional students: Educational and psychological procedures (6th ed.). Boston: Allyn and Bacon.

Woods, M.L., & Moe, A.J. (2003). Analytical Reading Inventory (6th ed.). Saddle River, NJ: Merrill.

Copyright of Psychology in the Schools is the property of Jossey-Bass, A Registered Trademark of Wiley Periodicals, Inc., A Wiley Company and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.