Accuracy of student performance while reading leveled books rated at their instructional level by a reading inventory

Matthew K. Burns [a,*], Sandra M. Pulles [b], Kathrin E. Maki [b], Rebecca Kanive [b], Jennifer Hodgson [b], Lori A. Helman [b], Jennifer J. McComas [b], June L. Preast [a]

[b] University of Missouri, United States
[b] University of Minnesota, United States

ARTICLE INFO

ABSTRACT

Identifying a student's instructional level is necessary to ensure that students are appropriately challenged in reading. Informal reading inventories (IRIs) purport to assess the highest reading level at which a student can accurately decode and comprehend text. However, the use of IRIs in determining a student's instructional level has been questioned because of a lack of research. The current study examined the percentage of words read correctly with 64 second- and third-grade students while reading from texts at their instructional level as determined by an IRI. Students read for 1 min from three leveled texts that corresponded to their instructional level as measured by an IRI, and the percentage of words read correctly was recorded. The percentage read correctly correlated across the three books from $r = .47$ to $r = .68$ and instructional level categories correlated from $tau = .59$ to $tau = .65$. Percent agreement calculations showed that the categorical scores (frustration, instructional, and independent) for the three readings agreed approximately 67% to 70% of the time, which resulted in a kappa estimate of less than .50. Kappa coefficients of .70 are considered strong indicators of agreement. Moreover, more than half of the students with the lowest reading skills read at a frustration level when attempting to read books rated at their instructional level by an IRI. The current study questions how reliably and accurately IRIs identify students' instructional level for reading.

1. Introduction

As Allington (2002) stated, "you can't learn much from books you can't read" (p. 16). There is an entire industry in education based on providing students appropriately challenging reading material. As a result, the term 'instructional level' is one of the most frequently used in education today and generally refers to providing an appropriate level of challenge in which students are sufficiently engaged but not bored or frustrated (Gravois & Gickling, 2008). If the learning task is too difficult, then the students will be frustrated, but tasks that are too easy could result in student boredom. Thus, providing an appropriate level of challenge is one feature of effective academic interventions (Burns, VanDerHeyden, & Zaslofsky, 2014; Vaughn, Gersten, & Chard, 2000), and should be part of any assessment-to-intervention model (Daly, Witt, Martens, & Dool, 1997), but there is considerable variability in methods to determine an instructional level.

Betts coined the term 'instructional level' in 1946 to describe the appropriate level of challenge for reading when he anecdotally noted that children generally read better when they correctly read about 95% of the words. Betts simultaneously began to develop

---

* Corresponding author at: 109 Hill Hall, Columbia, MO 65211, United States. Tel.: +1 573 882 8192.
E-mail address: burnsmi@missouri.edu (M.K. Burns).
ACTION EDITOR: Renee Hawkins.

assessment techniques to better understand student reading based on the percentage of words that a student could accurately read (Pikulski, 1974), which evolved into what is today referred to as an informal reading inventory (IRI). IRIs are designed to assess the highest reading level at which a child can accurately read the words and comprehend the text (Nilsson, 2013), and are commonly used in schools (Mellard, McKnight, & Woods, 2009; Paris, 2002; Paris, Paris, & Carpenter, 2002).

IRIs typically involve having students read aloud from vocabulary lists or passages written to represent specific grade or developmental levels while an assessor follows along to identify errors in reading (Nilsson, 2013). The highest level at which a student can read 90% to 95% of words while demonstrating sufficient comprehension and fluency, as judged by the assessor, is identified as the student's instructional level. IRIs have long been used as a diagnostic tool to determine student instructional needs (Nilsson, 2013), and there are currently dozens of IRIs that are published by test and curriculum publishers, with some being in their 10th edition (e.g., Johns, 2010). However, critiques emerged shortly after IRIs were first developed because research found the potential for considerable measurement error in the data (Lowell, 1970; Pikulski, 1974). Recently, scholars have questioned the reliability of data obtained from IRIs because reliability coefficients were not consistently reported (Nilsson, 2008). An evaluation of nine recently published IRIs found that only four included reliability data, and of those approximately half of the coefficients were at or below .80 (Spector, 2005). Therefore, more research is needed to evaluate decisions made with IRI data.

The Fountas and Pinnell Benchmark Assessment System: Second Edition (BAS; Heinemann, 2012) is a recently published IRI that is commonly used in schools. Published test–retest reliability between fiction and informational texts was .97 (Heinemann, 2012), but independent research reported test–retest reliability of .86 (Klingbeil, McComas, Burns, & Helman, 2015). Convergent criterion-validity estimates were $r = .94$ with reading scores obtained with texts from Reading Recovery, $r = .44$ with Degrees of Reading Power assessment (Touchstone Applied Science Associates, 1995), and $r = .69$ with the Slosson Oral Reading Test — Revised (Slosson & Nicholson, 2002; Heinemann, 2012). There has been limited independent research of the BAS. Parker et al. (2015) examined the diagnostic accuracy of BAS data in identifying struggling readers, using the Measures of Academic Progress for Reading (Northwest Evaluation Association, 2003) as the criterion measure, with over 900 second and third graders. Data from the BAS identified students as struggling or proficient consistently with the criterion only 54% of the time, which was roughly equal to chance (Parker et al., 2015). Thus, the BAS seems to be more promising than previously published IRIs that frequently do not provide estimates of reliability and validity (Spector, 2005), but independent research questions the utility of the data and suggests that more research is needed.

Although there are numerous published IRIs, reading teachers initially relied on data taken from student instructional materials rather than commercially prepared samplings of multiple curricula (Pikulski, 1974). Gickling and Armstrong (1978) operationally defined Bett's (1946) concept of an instructional level for reading as material in which the student could accurately read 93% to 97% of the words. The assessments in the Gickling and Armstrong (1978) study were taken directly from the materials used for reading instruction and provided the basis for what became known as curriculum-based assessment for instructional design (CBA-ID, Coulter & Coulter, 1990; Gickling & Havertape, 1981). In CBA-ID, students read orally from their learning materials (e.g., reading basal) for three 1-minute samples, and the assessor records the number of words read correctly and the total number of words. Next, the number of words read correctly is divided by the total number of words and multiplied by 100 to get a percentage score, which is then compared to the instructional level criterion of 93% to 97%. If the student read fewer than 93% of the words correctly, that would represent a frustration level, and more than 97% correct words would indicate a student's independent level. If a student read at the frustration level (less than 93% of the words correct), then the material was probably inappropriate for instruction. Selecting material in which students read 93% to 97% correct led to increased task completion, task comprehension, and time on task during reading instruction (Gickling & Armstrong, 1978; Treptow, Burns, & McComas, 2007). Alternatively, teachers could preteach words from the curriculum until the student could read 93% of the words correctly, which leads to increased student learning (Burns, 2007).

There is considerable research supporting the use of CBA-ID to make instructional decisions. As stated above, selecting material in which students read 93% to 97% of the words correctly increased task completion and comprehension, and time on task (Gickling & Armstrong, 1978; Treptow et al., 2007), and using CBA-ID data to modify instruction accelerated student learning (Burns, 2002, 2007; Shapiro & Ager, 1992). Moreover, previous research regarding the psychometric properties of assessing the instructional level within CBA-ID found that the approach resulted in interscorer reliability coefficients that ranged from .89 to .99, internal consistency coefficients of .87 to .96, alternate form-reliability estimates from .80 to .86, and test–retest coefficients, with a 2-week test–retest interval, that ranged from .82 to .96 (Burns, 2001; Burns, Tucker, Frame, Foley, & Hauser, 2000).

1.1. Purpose

Teachers seem to rely heavily on assessments of the instructional level to design instruction, select reading material for students, and assign guided reading groups (Nilsson, 2008). Moreover, assessing if the interaction between task demand and student skills represents an instructional level could be an important variable in analyzing student problems (Roberts, Marshall, Nelson, & Albers, 2001), and the difficulty of material to which students are expected to respond is an important factor to consider in designing interventions (Daly et al., 1997). However, there is little research regarding decisions made with IRIs and none that compares the decision to data from CBA-ID, for which there is a stronger research base.

Assessment research in school psychology has historically relied on correlations between similar measures (Burns, 2011), often referred to as criterion-validity (American Educational Research Association [AERA], American Psychological Association [APA], & National Council for Measurement in Education [NCME], 1999). However, relying on correlations between two similar measures to evaluate assessment data is a "weak program" (p. 326) that results in conceptual circularity (Kane, 2001), and does not adequately capture the concept of validity (AERA, APA, NCME, 1999). Validity evidence should focus on a science of diagnosis that researches meaningful decision thresholds, the diagnostic accuracy associated with those thresholds (Swets, Dawes, & Monahan, 2000), and the reliability of