Journal of School Psychology 53 (2015) 437–445

Contents lists available at ScienceDirect

Journal of School Psychology

journal homepage: www.elsevier.com/locate/jschpsyc

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

Article history:
Received 28 February 2014
Received in revised form 22 September 2015
Accepted 23 September 2015
Available online xxxx

Keywords:
Informal reading inventory
Instructional level
CBA-ID

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.

http://dx.doi.org/10.1016/j.jsp.2015.09.003
0022-4405/

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

decisions made with the data (AERA et al., 1999). Therefore, the current study was conducted to be the first in a line of inquiry to examine how well children read from books designed to represent their instructional level, as determined by IRIs, by studying the consistency with which students read passages purportedly written at the same level and how accurately they read passages that were purportedly at their instructional level. The following research questions guided the study, (a) How consistent are the instructional level estimates based on accuracy while reading three books rated to be at the same difficulty level by an IRI? (b) To what extent do students accurately read books that are rated at their instructional level using IRI data? (c) How do reading skills affect the accuracy with which students read books identified to be at their instructional level with IRI data?

2. Method

2.1. Participants and setting

The study participants were 64 second- and third-grade students attending one elementary school in Minnesota. A total of 28 (43.8%) of the students were second grade and 36 (56.3%) were in third grade, and 33 (51.6%) were female and 31 (48.4%) were male. Finally, 39 (60.9%) of the students were White, 12 (18.8%) were African-American, 8 (12.5%) were Hispanic, and 5 (7.8%) were Asian-American. None of the students received special education services for reading.

The school that the students attended served 630 students in kindergarten through eighth grade in an urban setting. A total of 55% of the students were White and 45% were from an ethnic minority background. Approximately 40% of the students who attended the school were eligible for the federal free or reduced price lunch. The school used a balanced literacy approach to reading instruction that included guided reading groups based on the Fountas and Pinnell (1996) reading program.

2.2. Measures

2.2.1. Oral reading fluency

Students attending the participating school completed standardized oral reading fluency (ORF) benchmark assessments three times during the academic year (September, January, and May). Each student read from grade-level probes from the AIMSweb (Pearson, 2008) assessment system for 1 min while an assessor followed along and recorded the number of words read correctly and the number of errors. Three assessments were conducted for each student at each benchmark assessment and the median score was recorded. ORF assessments have been shown to have strong alternate-form reliability of .95 (Good, Kaminski, Smith, & Bratten, 2001) and test–retest reliability of .96 (Catts, Petscher, Schatschneider, Bridges, & Mendoza, 2009). Moreover, ORF scores are moderately correlated with several standardized tests of reading achievement (Deno, Mirkin, & Chiang, 1982; Reschly, Busch, Betts, Deno, & Long, 2009) and state assessments (e.g., Hintze & Silberglitt, 2005; McGlinchey, & Hixon, 2004; Shapiro, Keller, Lutz, Santoro, & Hintze, 2006; Stage & Jacobsen, 2001). The data for the current study consisted of the median words read correctly per minute (WRCM) from the three 1-minute ORF assessments conducted during the May benchmark assessment.

The ORF data were used to determine reading skills for the students at the time of the study. Thus, scores of 69 WRCM or lower for second grade and 84 WRCM or lower for third grade were identified as low readers because those scores fell at or below the 25th percentile. Scores that met or exceeded 120 WRCM for second grade and 140 for third grade represented the 75th percentile or above and were classified as high readers. Finally, scores that fell between 69 and 120 for second grade, and 84 and 140 for third grade were classified as average readers.

2.2.2. Fountas & Pinnell Benchmark Assessment System

The Fountas & Pinnell Benchmark Assessment System: Second Edition (BAS; Heinemann, 2012) is designed to measure accuracy, oral fluency, and reading comprehension. There are two assessment systems, one for students in grades K-2 and another for students in grades 3–8, System 1 and System 2 respectively. The measure is intended for use in tiered student support systems and designed for triennial screening (Heinemann, 2012). Scores for the BAS are letter designations that represent an instructional level for students for books rated by the system. For example, a student who scored an instructional level of M would read from books rated as an M because those books would represent an instructional level.

In order to determine the level of book at which to start the BAS, the administrator uses a word list provided by the BAS (Heinemann, 2012). Scores on the word list indicate where to begin the assessment. Each student is asked to read a leveled passage that is not timed and to answer comprehension questions to determine if the passage represents a hard level, instructional level, or independent level. The assessment continues by asking the student to read progressively more difficult levels until the student's accuracy and comprehension fall within the hard level range. The highest level at which the student scored within the instructional level range is identified as the student's instructional level. All test administrators followed the standardized procedures, which meant that the test began with the word-list assessment, the passages were untimed, and the assessment continued until the accuracy and comprehension scores fell within the hard range. The district required all teachers to follow the standardized practice and each was trained to do so before beginning the assessments.

Published test–retest reliability ranged from .86 (Klingbeil et al., 2015) to .97 (Heinemann, 2012), and convergent validity estimates with other measures of reading ranged from $r = .44$ to $r = .94$ (Heinemann, 2012). The BAS was administered one-on-one by the student's classroom teacher during the May assessment period, which was approximately 1 week prior the current study. Second-grade students completed System 1 and third-grade students completed System 2. The data for the current study consisted

of the letter that corresponded to the instructional level, which ranged from B ($n = 1$, 1.6%) to Z ($n = 3$, 4.7%) with Q ($n = 7$, 10.9%) being the most frequently observed score. Each student's score was provided by the classroom teacher.

2.2.3. CBA-ID

The accuracy with which students read books rated to be at their instructional level by the BAS was assessed with CBA-ID. The students were presented with a book that was rated at their letter level (e.g., an M book for a student whose instructional level is rated as M) based on BAS guidelines, and they were asked to orally read the text using procedures outlined by Burns and Parker (2014), which are quite similar to procedures used for other curriculum-based approaches such as curriculum-based measurement. The instruction involved saying to the student: "Start right here and please read this out loud so that I can hear you. Do your very best reading. If you come to a word that you don't know, I'll tell it to you. Keep reading until I tell you to stop. Ready? Begin."

Students then read orally for 1 min. Words read correctly within 2 s were considered correct and those not read correctly within 2 s (e.g., omissions, word provided did not match written word, or correct reading after 2 s) were counted as errors. Students were told the correct pronunciation of words after 2 s expired without a correct response. Two seconds was used as the latency criterion to be consistent with CBA-ID procedures (Burns & Parker, 2014). In the event the student read all of the words on the given page in less than 1 min, the student was instructed to turn the page and continue reading until 1 min expired. All of the leveled books provided to the students contained pictures with each page of text, and none of the students read to the end of the book in less than 1 min.

After reading for 1 min, the assessor then counted the number of words read correctly, divided by the total number of words (words correctly read plus number of errors), and multiplied by 100. Scores that fell within the criterion of 93% to 97% correct were identified as an instructional level, those that fell below 93% were identified as a frustration level, and those that were above 97% represented an independent level. This process was repeated two additional times with two additional books rated at students' identified level (total of three).

2.3. Procedures

ORF and BAS data were collected in separate assessment sessions by school personnel approximately 1 week before the study occurred. The additional data collection for the study occurred in one session in May. Students were taken to a quiet place within the school but outside of the classroom. Three books were randomly selected from the box of books that matched the student's instructional level according to BAS guidelines (e.g., three books rated as an M for a student with an instructional level of M according to the BAS), which was provided by the student's classroom teacher. Books were randomly selected by reaching into a bin with leveled books in them (e.g., a bin containing only books rated as an M) and selecting one. This procedure was followed three times.

Next, one page was selected from each book with a random number generator that represented the page numbers of the book. The student was asked to read orally for 1 min starting on the selected page as the assessor followed along with a second copy of the book. Words read correctly and errors were noted to compute a percentage of words read correctly for each timed reading from the three books. Agreement between the BAS estimate and student performance while reading the books was computed by determining the percentage of students whose reading from the leveled book represented an instructional level (i.e., correctly read 93% to 97% of the words). The entire assessment session, including all three books, for each student required approximately 5 min to complete.

2.4. Interobserver agreement

A total of 23 (36%) students were observed by a second person while orally reading the books in order to compute interobserver agreement (IOA). The number of words that each observer consistently rated as accurately or not accurately read was divided by the total number of words read during the assessment and multiplied by 100. The resulting IOA was 100% across the 23 observations.

2.5. Analyses

The association between percentages of words read correctly across the three probes for each student was evaluated with two types of correlations. First, a Pearson r was used to correlate the percentages as continuous data. Second, the data were converted to categorical of frustration (<93%), instructional (93% to 97%) and independent (>97%) and correlated with a Kendall's tau (τ). Finally, percent agreement and Kappa coefficients were computed as well.

Table 1
Descriptive data for the oral reading fluency benchmark assessment and the three reading assessments during the study.

	Mean (*SD*)	Reading Level		
		Frustration *n* (%)	Instructional *n* (%)	Independent *n* (%)
Spring benchmark oral reading fluency	128.08 (48.71)	NA	NA	NA
Reading 1 accuracy	96.7% (3.27%)	8 (12.5%)	23 (35.9%)	33 (51.6%)
Reading 2 accuracy	96.4% (4.91%)	10 (15.6%)	15 (23.4%)	39 (60.9%)
Reading 3 accuracy	96.1% (5.9%)	11 (17.2%)	17 (26.6%)	36 (56.3%)
Median accuracy	96.7% (3.6%)	10 (15.6%)	18 (28.1%)	36 (56.3%)

Note. Accuracy is based on percentage of words read correctly.

The second research question was evaluated by presenting the frequency with which students read at an instructional level, according to CBA-ID procedures, when reading from a book that was rated at their instructional level. The data were further analyzed with a one-way analysis of variance (ANOVA) in which the BAS reading level was grouped into one of five groups with four reading levels in each to represent the independent variable, and the accuracy with which the student read the instructional level passage was the dependent variable. The third research question was descriptive in nature and mean percentages were reported.

3. Results

Before addressing the research questions, the data were examined for grade-level differences in Spring Benchmark ORF and median percent accurate. The mean ORF from the Spring Benchmark for the second grade students was 130.50 ($SD = 36.68$) words read correctly and it was 126.19 ($SD = 56.79$) for third grade. The difference between the two was not significant $t(62) = .35$, $p = .73$ and the effect was small $d = .09$. The mean percentage of words read correctly for the median of the three reading assessments was 96.68% ($SD = 3.61\%$) for second grade and 96.71% ($SD = 3.66\%$) for third grade, which resulted in a nonsignificant $t(62) = .04$, $p = .97$ and negligible $d = .01$ effect. Thus, data from the two grade groups were collapsed for subsequent analyses. The descriptive data for the variables are presented in Table 1.

The first research question inquired about level of agreement between instructional level estimates obtained from reading three books using CBA-ID procedures in corresponding leveled books. As shown in Table 1, the mean percentage of words read correctly was fairly stable across the three assessments. However, the percentage of assessments that fell within the instructional level ranged from 23.4% to 35.9%. Thus, the data were correlated to further examine the relationship. As shown in Table 2, the percent read correctly correlated from $r = .47$ to $r = .68$. The instructional level categories correlated from $\tau = .59$ to $\tau = .65$, which suggested moderate associations. Finally, the percent agreement data presented in Table 3 suggested 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. Thus, these data suggest considerable variability in reading performance for these three readings taken from books rated at the same level of difficulty.

The second research question inquired about agreement between estimates of instructional level from a reading inventory and instructional level estimates from reading the corresponding leveled books. The mean accuracy of 96.7% ($SD = 3.6\%$) falls within the instructional level range of 93% to 97%. However, the frequency of categorical scores for the median percentage of words read correctly, reported in Table 1, indicated that only 28.1% of the median scores represented an instructional level when reading from a book rated to represent the instructional level determined by the reading inventory. The reading books were likely too difficult for 15.6% of the students and could have been too easy for 56.3%. The question was further examined with a one-way ANOVA using the BAS reading level group as the independent variable and the median percentage of words read correctly while reading the instructional level book as the dependent variable. The mean score for students at each reading level is presented in Table 4. The scores were grouped arbitrarily into five groups with four reading levels in each in order to conduct an ANOVA. The result of the analysis was significant with a large effect $F(4, 59) = 13.80$, $p < .001$, $\eta^2 = .48$. Students at the lower reading levels tended to read with less accuracy, even though they were reading from books that were rated at their instructional level.

The third research question addressed the effect that reading skills would have on the agreement between estimate of instructional level from a reading inventory and estimates from reading the corresponding leveled book. These data are presented in Table 5, which includes frequency estimates for the median percentage of words read correctly falling within frustration, instructional, and independent levels for students in the low, middle, and high reading skills groups. As shown in Table 5, students who read below the 25th percentile on the spring benchmark assessment read at the instructional level 41.7% of the time, but the books that were rated as an instructional level for these students represented a frustration level 58.3% of the time. Moreover, the instructional level estimates from the reading inventory seem to underestimate the reading skills of the middle and high students because the score fell within the independent level 71.4% and 67.7% of the time respectively.

4. Discussion

The current study found that students inconsistently read from books that were rated to be at the same level of difficulty. These data suggested that students were only somewhat consistent across reading samples, or the samples of the same level varied somewhat from one another in difficulty. Thus, the first research question more closely addresses consistency in reading accuracy scores obtained from multiple texts classified to be at the same reading level, than the BAS classification accuracy. Previous research also found that students performed differently on similar level texts because of the text structure, prior knowledge, the nature of the

Table 2
Correlations of accuracy (percentage of words read correctly) from three reading performance assessments.

	Reading 1 accuracy	Reading 2 accuracy	Reading 3 accuracy
Reading 1 accuracy		$r = .47^*$	$r = .61^*$
Reading 2 accuracy	$\tau = .67^*$		$r = .68^*$
Reading 3 accuracy	$\tau = .67^*$	$\tau = .59^*$	

Note. Correlations above the diagonal are Pearson product moment correlations using the percent read correctly and those below the diagonal are Kendall's tau (τ) with categorical scores of frustration, instructional, and independent.
* $p < .01$.

Table 3
Percent agreement and kappa from accuracy measures (percentage of words read correctly) from three reading performance assessments.

	Reading 1 accuracy	Reading 2 accuracy	Reading 3 accuracy
Reading 1 accuracy		70.3%	68.8%
Reading 2 accuracy	$k = .49^*$		67.2%
Reading 3 accuracy	$k = .47^*$	$k = .42^*$	

Note. Data are converted to categorical scores of frustration, instructional, and independent. Data above the diagonal are percentage agreement, and the estimates below the diagonal are kappa.
* $p < .01$.

assigned reading task, or how the reading level was defined (Calisir & Gurel, 2003; Christ, White, Ardoin, & Eckert, 2013; Hiebert & Pearson, 2010). However, there is no way to determine which of the potential sources of variability most affected the results in the current study without additional research.

The analysis for the second and third research questions found that only about one quarter of the time did the students read 93% to 97% of the words correctly when reading the book that was rated at their instructional level, and students who were struggling readers frequently failed to read at least 93% of the words correctly when they were reading from a book that was designated by an IRI to provide an appropriate level of difficulty. The accuracy with which students read books at their instructional level could question the assigned difficulty level of the reading material or the assessment's ability to identify a student's instructional level. Either way, these data question the validity of instructional decisions made with the IRI data. Validity evidence has historically relied on correlations between similar measures (Burns, 2011; Kane, 2001), but it should focus more on meaningful decision thresholds and the diagnostic accuracy associated with those thresholds (Swets et al., 2000). It seems that the reading level assigned by the assessment did not align with how well the student actually read text.

The current findings also could have some implications for assessment research because they suggest a need to caution against relying solely on criterion-related validity estimates to evaluate the validity of decisions made with assessment data. IRIs correlate well with other measures of reading (Fountas & Pinnell, 2007). However, previous research found that IRI data did not accurately identify students with reading problems (Parker et al., 2015) and the current data indicated that the IRI data did not identify reading material that students can successfully read. Screening reading skills of students (Heinemannn, 2012; Paris et al., 2002) and identifying instructional reading levels for students (Roe & Burns, 2010; Shanker & Cockrum, 2013) are the two purposes purported by IRI developers and users. Neither of the purported purposes could be adequately evaluated with correlations to other measures, and neither was supported by previous or current research.

Previous critiques of IRIs discussed poor or unreported reliability (Nilsson, 2008; Spector, 2005). For example, most IRIs do not report reliability by age or grade-level, but instead aggregate the data across multiple grades, which could artificially inflate the reliability estimate (Spector, 2005). Therefore, one potential reason the current data did not support the accuracy of IRI estimates could be poor reliability within individual grade levels. However, it is important to note that there were slight variations in accuracy metrics for each measure. The CBA-ID criterion for an instructional level is 93–97%, but 90–94% are used for levels A through K with the BAS, and 95%–97% for levels for the remaining levels. Although the criteria are largely consistent, they are different and could explain some of the variation in performance across the two measures. Although reliability is an important construct, issues of generalization are also of concern. One potential reason for the low match between instructional level estimates and reading performance on texts rated at that level could be poor generalizability. Data obtained from a test should represent the content that the test is intended to

Table 4
Mean accuracy for students in each benchmark assessment system reading level.

Reading level	*N*	Mean	*SD*	Level group	*N*	Mean	*SD*
B	1	83.33%	NA	1 (B–G)	4	89.07%	3.97%
D	1	90.91%	NA				
F	1	92.31%	NA				
G	1	89.74%	NA				
J	3	95.72%	4.17%	2 (J–N)	15	94.70%	4.00%
L	1	95.00%	NA				
M	5	94.84%	2.99%				
N	6	94.00%	5.41%				
O	1	97.85%	NA	3 (O–R)	18	97.21%	2.49%
P	6	96.64%	2.75%				
Q	7	97.91%	1.85%				
R	4	96.67%	3.68%				
S	4	98.94%	0.85%	4 (S–V)	18	98.41%	1.51%
T	6	98.09%	1.93%				
U	6	98.32%	1.77%				
V	2	98.55%	0.59%				
W	3	99.16%	0.78%	5 (W–Z)	9	98.98%	1.07%
X	1	100.00%	NA				
Y	2	98.02%	1.91%				
Z	3	99.11%	0.79%				

Table 5
Number and percentage of median percentage of word read correctly scores from three reading performance assessments that fell within the frustration, instructional, and independent level by skill group.

Group	Frustration *n* (%)	Instructional *n* (%)	Independent *n* (%)
Low — 25th percentile or less	7 (58%)	5 (41.7%)	0 (0.0%)
Middle — 26th To 75th percentile	2 (9.5%)	4 (19.0%)	15 (71.4%)
High — 76th percentile or higher	1 (3.2%)	9 (29.0%)	21 (67.7%)

sample (Salvia, Ysseldyke, & Bolt, 2010), but data from the IRIs used in this study may not generalize well to the universe of books that they were designed to represent. It would be difficult to design an assessment system that does generalize well to reading levels of authentic books because successful reading is dependent on background knowledge, vocabulary from the given text, and text structures (Calisir & Gurel, 2003; Cromley & Azevedo, 2007), which would be almost impossible to assess or take into account for an individual student.

Perhaps the best method to assure generalizability is to conduct the assessment within students' actual instructional material because that would dramatically reduce the assumptions of generalization. Pikulski (1974) suggested that "the primary strength of an IRI, that of close correspondence between the test material and the teaching material, is lost" (p. 142) when teachers use commercially prepared IRIs rather than conduct the analyses with the actual reading material. However, some level of standardization is needed to ensure a valid decision (Salvia et al., 2010), which can be accomplished by following CBA-ID procedures. Another potential reason that IRI estimates did not correlate well with instructional level estimates when reading the books could have been a mismatch between criteria. IRIs tend to include direct measures of comprehension in their estimates of instructional level and CBA-ID uses percentage of words read correctly as an indicator of comprehension. There is an almost linear association between percentage of words that students can read and reading comprehension (Schmitt, Jiang, & Grabe, 2011), and correctly reading words and comprehension tend to grow together with quite high correspondence (Torgesen, 2006). Moreover, decoding and word recognition ($r = .61$ and .60 respectively) correlated more strongly with a measure of reading comprehension than did measures of receptive vocabulary ($r = .44$), expressive vocabulary ($r = .36$) or depth of vocabulary knowledge ($r = .30$) among a group of fourth-grade students (Oullette, 2006), and accuracy together with rate of real word reading predicted reading comprehension among struggling readers in second grade (Berninger, Abbott, Vermeulen, & Fulton, 2006). Finally, although there are certainly children with poor comprehension who are accurate readers (Cain & Oakhill, 2006), there are very few who read with low accuracy and high comprehension (Meisinger, Bradley, Schwanenflugel, & Kuhn, 2010). However, reading comprehension involves language comprehension and vocabulary in addition to word recognition (Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001; Snowling, 2005). Perhaps if students were asked to answer comprehension questions based on what they read, and those data contributed to the instructional level estimates, different results might have been obtained. Future research should consider both accuracy and comprehension when studying how well children interact with specific texts. Perhaps the reciprocal nature of accurate word reading and comprehension allows for the percentage of words read correctly to serve as an indication of comprehension as well, but that is a hypothesis in need of research.

4.1. Limitations

Although the findings of this study show students' accuracy (as measured by the percentage of words read correctly) on IRI books was inconsistent, the results should be interpreted within the context of their limitations. First, the mean for words read correctly on the spring oral reading fluency benchmark was 128.08, which is above the 2nd and 3rd grade benchmarks. Moreover, the median percentage of words read correctly across all three passages was 96.7%. Thus, most of the students were reading at a proficient level and did not adequately represent all readers. There were only 12 students who scored below the 25th percentile on the ORF measure, which suggests that those data should likely be used only as the basis for future research. Second, there was no control over prior student exposure to the texts used in this study. Although books were chosen at random, students may have read the book previously, impacting their performance during the study. Third, the study did not take the type of text (i.e., fiction versus informational) into account. Background knowledge of book content may have impacted student reading performance. Fourth, there was a limited sample of words (1 min) read from each leveled book and the overall text may have been more difficult or easier depending on the section of text read. Previous research found that the percentage of words read correctly was consistent when assessed with different sections of a book (Burns et al., 2000), but the potential implications of using only a 1-minute timing is unknown. Moreover, reliability data were collected for the ORF measure, but not the BAS. Although some aspects of the BAS are straightforward to score, and the district trained and required standardized administration, the extent to which the teachers actually followed the required procedures is unknown. Future researchers should evaluate the reliability of administration procedures, scoring procedures, and resulting data for the BAS and other IRIs. Fifth, the study used the BAS, which is only one of many published IRIs. It is unknown how well the current data generalize to other IRIs, which suggests an area for future research. Finally, the estimate of the instructional level was based on the percentage of words read correctly, but skill levels were based on ORF, which could partially explain the differences noted in Table 5.

4.2. Conclusion

Although IRIs are frequently used in schools to identify student instructional level, the current study questions how well IRIs identify students' reading instructional level. Moreover, IRIs are used to drive reading instruction, even though the current research suggests that instruction based on IRI identified instructional level may not be matched to student need. Research shows that instructional match results in improved student outcomes (Burns, 2007; Gickling & Armstrong, 1978; Treptow et al., 2007); thus it is necessary to use assessment measures that result in reliable data and valid decisions regarding student performance. Given the importance of developing reading skills, additional research concerning the use of IRIs for determining instructional level is warranted.

References

Allington, R. L. (2002). You can't learn much from books you can't read. *Leadership*, *60*(3), 16–19.

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

Berninger, V. W., Abbott, R. D., Vermeulen, K., & Fulton, C. M. (2006). Paths to reading comprehension in at-risk second-grade readers. *Journal of Learning Disabilities*, *39*(4), 334–351.

Betts, E. A. (1946). *Foundations of reading instruction, with emphasis on differentiated guidance*. Oxford: American Book Co.

Burns, M. K. (2001). Measuring sight-word acquisition and retention rates with curriculum-based assessment. *Journal of Psychoeducational Assessment*, *19*(2), 148–157.

Burns, M. K. (2002). Comprehensive system of assessment to intervention using curriculum-based assessments. *Intervention in School and Clinic*, *38*(1), 8–13.

Burns, M. K. (2011). School psychology research: Combining ecological theory and prevention science. *School Psychology Review*, *40*, 132–139.

Burns, M. K., & Parker, D. (2014). *Curriculum-based assessment for instructional design: Using data to design instruction*. New York: Guilford.

Burns, M. K., Tucker, J. A., Frame, J., Foley, S., & Hauser, A. (2000). Interscorer, alternate-form, internal consistency, and test–retest reliability of Gickling's model of curriculum-based assessment for reading. *Journal of Psychoeducational Assessment*, *18*(4), 353–360.

Burns, M. K., VanDerHeyden, A. M., & Zaslofsky, A. F. (2014). Best practices in delivery intensive academic interventions. In A. Thomas, & J. Grimes (Eds.), *Best practices in school psychology* (pp. 129–142) (6th ed.). Bethesda, MD: National Association of School Psychologists.

Burns, M. K. (2007). Reading at the instructional level with children identified as learning disabled: Potential implications for response-to-intervention. *School Psychology Quarterly*, *22*, 297.

Cain, K., & Oakhill, J. (2006). Profiles of children with specific reading comprehension difficulties. *British Journal of Educational Psychology*, *76*, 683–696.

Calisir, F., & Gurel, Z. (2003). Influence of text structure and prior knowledge of the learner on reading comprehension, browsing and perceived control. *Computers in Human Behavior*, *19*, 135–145.

Catts, H. W., Petscher, Y., Schatschneider, C., Bridges, M. S., & Mendoza, K. (2009). Floor effects associated with universal screening and their impact on the early identification of reading disabilities. *Journal of Learning Disabilities*, *42*, 163–176.

Christ, T. J., White, M. J., Ardoin, S. P., & Eckert, T. L. (2013). Curriculum-based measurement of reading: Consistency and validity across best, fastest, and question reading conditions. *School Psychology Review*, *42*, 415–436.

Coulter, W. A., & Coulter, E. M. (1990). Curriculum-based assessment for instructional design: Trainer's manual. Unpublished training manual available from Directions and Resources, P.O. Box 57113, New Orleans, LA 70157.

Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, *99*, 311.

Daly, E. J., Witt, J. C., Martens, B. K., & Dool, E. J. (1997). A model for conducting a functional analysis of academic performance problems. *School Psychology Review*, *26*, 554–574.

Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children*, *49*, 36–45.

Fountas, L. C., & Pinnell, G. S. (1996). *Good reading: Good first teaching for all children*. Portsmouth, NH: Heinemann.

Fountas, I. C., & Pinnell, G. S. (2007). *Fountas and Pinnell benchmark assessment system 2*. Portsmouth, NH: Heinemann.

Gickling, E., & Havertape, S. (1981). *Non-test based assessment training manual from National School Psychology Inservice Training Network*.

Gickling, E. E., & Armstrong, D. L. (1978). Levels of instructional difficulty as related to on-task behavior, task completion, and comprehension. *Journal of Learning Disabilities*, *11*(9), 559–566.

Good, R. H., Kaminski, R. A., Smith, S., & Bratten, J. (2001). Technical adequacy of second grade DIBELS oral reading fluency passages. *World Health Organization Technical Report Series*, *8*.

Gravois, T. A., & Gickling, E. E. (2008). Best practices in instructional assessment. In A. Thomas, & J. Grimes (Eds.), *Best practices in school psychology V. 2*. (pp. 503–518). Bethesda, MD: National Association of School Psychologists.

Heinemannn (2012). *Fountas and Pinnell*. Portsmouth, NH: author (Retrieved 12/20/2012 from http://www.Heinemannn.com/fountasandpinnell/default.aspx).

Hiebert, E. H., & Pearson, P. D. (2010). *An examination of current text difficulty indices with early reading texts*. San Francisco, CA: TextProject, Inc.

Hintze, J. M., & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review*, *34*(3), 372.

Johns, J. L. (2010). *Basic reading inventory* (10th ed.). Dubuque, IA: Kendall/Hunt Publishing.

Kane, M. T. (2001). Concerns in validity theory. *Journal of Educational Measurement*, *38*, 319–342.

Klingbeil, D., McComas, J. J., Burns, M. K., & Helman, L. (2015). Comparison of predictive validity and diagnostic accuracy of screening measures of reading skills. *Psychology in the Schools* (in press).

Lowell, R. E. (1970). Problems in identifying reading levels with informal reading inventories. In W. Durr (Ed.), *Reading difficulties: Diagnosis, correction and remediation*. Newark, DE: International Reading Association.

McGlinchey, M. T., & Hixson, M. D. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review*, *33*, 193–203.

Meisinger, E. B., Bradley, B. A., Schwanenflugel, P. J., & Kuhn, M. R. (2010). Teacher's perceptions of word callers and related literacy concepts. *School Psychology Review*, *39*, 54–68.

Mellard, D., McKnight, M., & Woods, K. (2009). Response to intervention screening and progress-monitoring practices in 41 local schools. *Learning Disabilities Research and Practice*, *24*, 186–195. http://dx.doi.org/10.1111/j.1540-5826.2009.00292.x.

Nilsson, N. L. (2008). A critical analysis of eight informal reading inventories. *The Reading Teacher*, *61*, 526–536.

Nilsson, N. L. (2013). The reliability of informal reading inventories: What has changed? *Reading and Writing Quarterly*, *29*, 208–230. http://dx.doi.org/10.1080/10573569.2013.789779.

Northwest Evaluation Association (2003). *Measures of academic progress for reading*. Lake Oswego, OR: Author.

Oullette, G. P. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology*, *98*, 554–566.

Paris, S. G., Paris, A. H., & Carpenter, R. D. (2002). Effective practices for assessing young readers. In B. Taylor, & P. D. Pearson (Eds.), *Teaching reading: Effective schools, accomplished teachers* (pp. 141–160). Mahwah, NJ: Erlbaum.

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

Parker, D. C., Zaslofsky, A. F., Burns, M. K., Kanive, R., Hodgson, J., Scholin, S. E., et al. (2015). A brief report of the diagnostic accuracy of oral reading fluency and reading inventory levels for reading failure risk among second and third grade students. *Reading and Writing Quarterly* (in press).
Pearson (2008). *Aimsweb.* New York, NY: author.
Pikulski, J. (1974). A critical review: Informal reading inventories. *The Reading Teacher*, *28*, 141–151.
Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest*, *2*, 31–74.
Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology*, *47*, 427–469.
Roberts, M. L., Marshall, J., Nelson, J. R., & Albers, C. A. (2001). Curriculum-based assessment procedures embedded within functional behavioral assessments: Identifying escape-motivated behaviors in a general education classroom. *School Psychology Review*, *30*, 264–277.
Roe, B., & Burns, P. C. (2010). *Informal reading inventory: Preprimer to twelfth grade* (8th ed.). Belmont, CA: Wadsworth.
Salvia, J., Ysseldyke, J., & Bolt, S. (2010). *Assessment in special and inclusive education.* Belmont, CA.: Wadsworth.
Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*, 26–43.
Shanker, J. L., & Cockrum, W. (2013). *Ekwall/Shanker reading inventory* (6th ed.). Bloomington, MN: Pearson.
Shapiro, E. S., & Ager, C. (1992). Assessment of special education students in regular education programs: Linking assessment to instruction. *The Elementary School Journal*, *92*, 283–296.
Shapiro, E. S., Keller, M. A., Lutz, J. G., Santoro, L. E., & Hintze, J. M. (2006). Curriculum-Based Measures and Performance on State Assessment and Standardized Tests Reading and Math Performance in Pennsylvania. *Journal of Psychoeducational Assessment*, *24*(1), 19–35.
Slosson, R. L., & Nicholson, C. L. (2002). *Slosson Oral Reading Test—Revised.* East Aurora, NY: Slosson Educational Publications.
Snowling, M. J. (2005). Literacy outcomes for children with oral language impairments: Developmental interactions between language skills and learning to read. In H. W. Catts, & A. G. Kamhi (Eds.), *The connections between language and reading disabilities* (pp. 55–75). Mahwah, NJ: Erlbaum.
Spector, J. E. (2005). How reliable are informal reading inventories? *Psychology in the Schools*, *42*, 593–603. http://dx.doi.org/10.1002/pits.20104.
Stage, S. A., & Jacobsen, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review*, *30*, 407–419.
Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1–26.
Torgesen, J. K. (2006). Recent Discoveries from Research on Remedial Interventions for Children with Dyslexia. In M. Snowling, & C. Hulme (Eds.), *The science of reading: A handbook.* Oxford: Blackwell Publishers.
Touchstone Applied Science Associates (1995). *Degrees of Reading Power – Revised.* Minneapolis, MN: Author.
Treptow, M. A., Burns, M. K., & McComas, J. J. (2007). Reading at the frustration, instructional, and independent levels: The effects on students' reading comprehension and time on task. *School Psychology Review*, *36*, 159–166.
Vaughn, S., Gersten, R., & Chard, D. (2000). The underlying message in LD intervention research. *Council for Exceptional Children*, *67*, 99–114.