

# **Istation's Indicators of Progress**

## **Oral Reading Fluency**

### **Technical Report**

Istation

Dallas, Texas

Copyright 2020

## Introduction

When the National Reading Panel convened in the late 1990s, its members conducted extensive research on reading instruction and what works best to help students learn to read. The panel identified five areas of instruction, known as the “Big Five,” that are critical for the development of reading skills: phonemic awareness, phonics, vocabulary, text comprehension, and fluency (National Reading Panel, 2000). Istation’s Indicators of Progress Early Reading (ISIP™ ER) assessment was built around the Big Five and includes subtests that address each area. Istation is now introducing an automated oral reading fluency measure, designed to assess this critical skill in young students. It is a sophisticated, web-delivered measure of fluency that helps teachers assess oral reading fluency in less time. Istation’s Oral Reading Fluency assessment (ISIP ORF) complements the ISIP ER by providing an oral fluency measure that students in grades 1-3 can take for progress monitoring.

## Oral Reading Fluency

The ability to read fluently helps students make the developmental leap from *learning to read* to *reading to learn* faster and with less effort. Oral Reading Fluency (ORF), which is how accurately and quickly a student can read aloud, is predictive of overall reading success. ORF is a durable and stable measure, and research demonstrates that students grow in fluency as they progress in reading (Hasbrouck & Tindal, 2017a.). When students are able to read fluently, they remove the cognitive load of trying to sound out words and can focus on understanding what they read. Students grow in words correct per minute (WCPM) more rapidly in the younger grades, and growth continues throughout the elementary years (Hasbrouck & Tindal, 2006). Students with disabilities have slower growth in ORF (Wanzek, Al Otaiba, & Petscher, 2014).

ORF is predictive of success in reading comprehension (Burns et al., 2011; Eason et al., 2013; Trainin, Hiebert, & Wilson, 2015; Valencia et al., 2010) and for diverse students including African Americans (Hintze et al., 2002), American Indians and low socioeconomic whites (Pearce & Gayle, 2009), and English language learners (Quirk & Beem, 2012).

ISIP ORF has an automated scoring feature that uses the latest voice recognition technology from Boulder Learning Inc. (BLI), known as Fluent Oral Reading Assessment, or FLORA (Bolaños et al., 2013). Researchers at BLI developed FLORA to accurately assess word accuracy, reading rate, and expressiveness with high rates of agreement with human scorers.

This technology provides teachers an engaging and time-saving alternative to their normal routine of giving an oral reading assessment. Intended to give teachers more time to spend in the classroom teaching and less time assessing, ISIP ORF focuses on providing a fun and engaging activity for students to use and the flexibility for teachers to review the recordings when it is convenient for them. The passages are presented to students in a randomized fashion, and if students take the assessment once a month for progress monitoring, they will not see the same passage twice.

### **Istation Goals and Design Objectives for ISIP ORF**

ISIP ORF is designed for English-speaking students in grades K-3, and norms are provided for the winter and spring semesters for first grade students and for fall, winter, and spring for students in second and third grade. Norms are not provided for kindergarten students or students in the fall of first grade, as many of these students are not yet reading well enough to take an oral reading fluency measure. However, we provide passages for students at these young ages in the event that teachers and administrators want to start tracking fluency as soon as students can read.

The ISIP ORF application is fun and engaging for students, encouraging them to do their best on the assessment. When they enter the app, ISIP ORF puts them in the role of a voice actor in the Istation Recording Studio to record their passages, similar to what actors do in animated movies, television, and video games. The sound engineers, Dakota or Josey, introduce the task and provide directions to the students (see Figure 1). Each grade has grade-leveled passages that are a mixture of fiction and nonfiction, which the ISIP ORF system assigns randomly to students across the classroom and across months.



*Figure 1. The sound engineers introduce the task to students as voice talent.*

Teachers and administrators have the option of giving the assessment monthly to generate a running record, or it can be used less often for benchmarking or progress monitoring. ISIP ORF does not contribute to an overall ISIP ER score.

The ISIP Oral Reading Fluency assessment provides schools with the following features:

- grade-leveled passages that students read aloud and record digitally in the Istation Recording Studio;

- manual teacher scoring and automated scoring options;
- audio playback features that allow teachers to play, advance, pause, and rewind student recordings;
- online access to an archive of student recordings available to review at any time;
- scoring that analyzes oral reading proficiency and measures accuracy and fluency;
- note-taking features for teachers to document observations and remarks; and
- real-time reports for progress monitoring and identifying which students require more instruction.

### **Test Design and Development**

Research has long established the correlations between oral reading fluency and reading comprehension for elementary age students, particularly in grades 1-3 (Petscher & Kim, 2011). Theories that date to the 1970s and '80s posited that automaticity in reading would allow cognitive processes to focus on comprehension rather than decoding. This theory has been substantiated many times over by leading researchers and was accepted by the National Research Council's Committee on the Prevention of Reading Difficulties in Young Children (National Reading Council, 1998). As such, oral reading fluency assessments have been used for many years in elementary school settings as a screener to identify potential risks to future reading success as well as a proxy progress-monitoring tool for comprehension (Petscher & Kim, 2011).

Additional research suggests that oral reading rate is influenced by oral language, and vocabulary in particular makes significant contributions to oral reading rate (Eason et al., 2013). Oral reading fluency may predict reading success because as children automatize lower level skills, such as phonological processing and alphabetic decoding, they free their working memory to focus on more complex cognitive tasks, such as reading comprehension. This underscores the

intertwined relationship between semantic, orthographic, and phonological representations in word recognition as a component of reading comprehension. ISIP ORF also has strong predictive validity with academic assessments, such as the Iowa Test of Basic Skills (Schilling et al., 2007), particularly in reading comprehension (Valencia et al., 2010). Measuring oral reading fluency and then analyzing error types for either meaning, structure, or visual errors can provide educators with meaningful insights into which components of reading the student is struggling with. Because of the critical role oral reading fluency plays in reading comprehension, it is particularly important to assess oral reading fluency in elementary students at regular intervals.

### Creating the ISIP ORF Passages

Passages for Istation's oral reading fluency subtest were written by Brattle Publishing Group, LLC (BPG). BPG provides a variety of curriculum development services and has developed and edited hundreds of reading passages and assessment items for educational products and international curriculum materials. Using their team of experienced educational writers and editors and following the specifications set by Istation, BPG developed 120 leveled passages for kindergarten through third grade.

Development criteria given to BPG for each grade level included word count and grade-level readability based on the Flesch-Kincaid Grade Level scale, the ratio of fictional to informational passages, cultural sensitivity considerations, high-interest topics, and names and proper nouns that are decodable or follow regular orthographic patterns. For construct validity and to reduce form effects, the passages fall within the middle- to end-of-year range for each grade level on the Flesch-Kincaid Grade Level scale. Passages for grades 1-3 average around 60 words, while kindergarten passages average around 40 words per passage. To reflect the

[National Assessment of Educational Progress 2017 Reading Framework](#)'s assessment data requirements of a 50/50 distribution of literary and informational text by grade 4, ISIP ORF passages increase from 25% informational text in kindergarten and grade 1 to 50% informational text by grade 3. Informational passages include procedural (how-to), expository, and persuasive texts with structures such as descriptive, sequential, and compare/contrast. For both fictional and informational passages, the content and structure are universal in nature and high interest, and they require little to no background knowledge prior to reading.

Each passage produced by BPG went through their editing and review process before being passed to Istation for an internal review process. Istation's subject matter experts in reading instruction reviewed the passages for potential gender, racial, cultural, and ethnic biases, stereotypes, and microaggressions, followed by a review by the editing team. Passages with any questionable biases or stereotypes were revised and sent through the editing and review process again. Additionally, consideration was given to possible biases or discrimination based on location. Each passage is considered fair for all children, regardless of their geographical region or whether they live in an urban, rural, or suburban environment.

### **Validation of the Passage Reading Levels**

To maintain construct validity and alleviate potential form effects, ISIP ORF uses leveled passages, which conform to the middle- to end-of-year range on the Flesch-Kincaid Grade Level scale for grades 1-3. Kindergarten passages measure in the beginning- to middle-of-year range on the same scale. Great care was taken during the editing process to keep the readability within these ranges while maintaining authenticity. Since the passages within each grade level remain in the specified range, the text complexity increases as students age into the next grade, not as they

are taking the assessment. For additional accuracy, passages were sent to MetaMetrics® to obtain official Lexile® text measures, with the expectation that the Lexile measures would reflect similar grade level ranges as the Flesch-Kincaid Grade Level scale. The passages with Lexile measures that fell outside of the expected grade level ranges were removed from the testing bank. Table 1 shows the passages’ readability ranges using the Flesch-Kincaid Grade Level scale and Lexile text measures for each grade level.

Table 1. Range of Grade Leveled Passages for Flesch-Kincaid and Lexile Levels

<i>Grade</i>	<i>Flesch-Kincaid</i>	<i>Lexile Level</i>
<i>Kindergarten</i>	0.0 – 0.5	120L – 450L
<i>Grade 1</i>	1.5 – 1.9	270L – 560L
<i>Grade 2</i>	2.5 – 2.9	310L – 630L
<i>Grade 3</i>	3.5 – 3.9	490L – 810L

### Scoring the ISIP ORF

ISIP ORF provides two different ways to score the passages that the student records: manual scoring and automated scoring based on voice recognition technology. Manual (or human) scoring is accessed through the Istation application.

### Human Scoring

The teacher, reading intervention specialist, or administrator logs in to their Istation account to access the individual passages that each student recorded and manually scores each recording. They can mark errors and self-corrections while listening to the audio recording the



student created. The scorer can also mark that the errors or self-corrections were a result of a visual, meaning, or syntactic challenge. Teachers can make notes for each line of text or for an overview of the entire recording. The teacher will also mark the last word read and when on the timeline it occurs. This allows for the most accurate words correct per minute (WCPM) score.

If a passage is recorded by a student who is clearly unable to read, is experiencing technical difficulties, or has some other issue that arose during the recording that makes the audio file unable to be scored, the scorer has the ability to mark the passage as *unscorable*.

### **Automated Scoring**

Automated scoring compares the student's recorded audio to trained acoustic models for the original passage. The software is able to detect all omissions, insertions, and errors that are in the student's recorded passage, and it calculates the number of correct words.

Two decades of research indicate that a student's reading can be accurately assessed using automated speech recognition technology (Bolaños et al., 2013). Istation uses the technology developed by BLI. Working with oral reading fluency experts Jan Hasbrouck and Gerald Tindal, BLI developed FLORA for use with young children to capture their voices accurately. Research demonstrates that FLORA is highly correlated with human scoring. In a study of Colorado elementary school students, correlations between FLORA and the human scorers were strong. The mean difference between FLORA and the human scorer was approximately 3.6 words, and the mean difference between the two human scorers was approximately 1.2 words. Using the Hasbrouck and Tindal (2006) national norms for oral reading fluency, the researchers next compared FLORA- and human-scored passages with percentile ranks. The inter-rater agreement between FLORA and the human scores was strong at

.89, meaning that FLORA was reliable for identifying students who need additional help with reading (Bolaños et al., 2013).

## Dialects in American English and FLORA

Dialect is an important consideration for using voice recognition technology. In American English, there are several different dialects across the country that follow historical migration and immigration patterns. Upon recommendations from BLI, Istation used the dialect regions outlined by the Atlas of North American English (Labov, Ash, & Boberg, 2006) and used the website provided by Aschmann (2018) as references for the dialect regions. We segmented public school districts into dialect regions by zip code to match the dialect regions found in Table 2.

Some local dialects are missing from this classification, such as Tidewater (Down East), the specific accents around cities such as Charleston, South Carolina; Savannah, Georgia; and Greater New Orleans. We have also not classified specific ethnic dialects such as Cajun-influenced English, Spanish-influenced English, and African American English. Dialects are not disappearing all together, but the local dialects are merging with larger regional dialects (Labov, Ash, & Boberg, 2006); therefore, we merged areas with these localized dialects with the more geographically broad dialects. We also merged classical southern with Inland or Lowland South, as in these areas there can be a mixture of dialects that people speak.

Dialectical variations can confound the software algorithms; therefore, it is important that a wide variety of voices be recorded in a dialect before the software reaches reliability. If the software is not trained within a dialectical area, words read with these dialectical variations may receive an inaccurate score. Therefore, during the pilot phase, Istation trained the software in

several dialect regions, and only those regions that had sufficient voice samples and high correlations with human scoring will be released for automated scoring.

Table 2. Dialects and Subdialects by State

<i>Regional Dialects</i>	<i>Subdialect</i>	<i>Entire States in Subdialect Region</i>	<i>Partial States in Subdialect Region</i>
<b>North Dialects</b>	Canadian		Alaska
	Northern New England	Massachusetts, Maine, New Hampshire, Vermont	
	Inland North	Michigan, Rhode Island, Connecticut	Iowa, Illinois, Indiana, New York, Pennsylvania, Wisconsin
	Western North		Iowa, Minnesota, Nebraska, South Dakota, Wisconsin
	Greater New York City		New York (City)
<b>Midland</b>	East Midland	New Jersey, Delaware, Ohio	Delaware, Maryland, Ohio, Pennsylvania
	Central Midland		Iowa, Illinois, Indiana, Kansas, Missouri
	West Midland		Nebraska, Oklahoma
<b>South</b>	Inland South	Washington, DC	Alabama, Arkansas, Georgia, Kentucky, Missouri, North Carolina, Oklahoma, Tennessee, Texas, Eastern New Mexico (Hobbs, Carlsbad areas)
	Classic South/ Lowland South	Mississippi, West Virginia, Virginia, South Carolina, Louisiana	North Carolina, Georgia, Alabama, Florida
	North Florida		North Florida
	South Florida		South Florida
<b>North Central</b>	Inland Central		Wisconsin, Michigan, South Dakota, Eastern Montana
	North Central	North Dakota	Eastern Montana, Northwest South Dakota
<b>West</b>	West	Washington, Oregon, California, Utah, Idaho, Wyoming, Colorado, Nevada, Arizona	Western Montana, Western South Dakota, Western Kansas, New Mexico

## **Testing and Validation of Istation ORF**

Testing and validation of ISIP ORF occurred in three phases. During the first phase, conducted from November 2018 through February 2019, we collected voice samples to train the FLORA system on the ORF passages and monitored the correlation between human and automated scoring for the passages. In the second phase, conducted from March through May 2019, we evaluated item difficulties and conducted inter-rater reliability studies between automated and human scoring and between two human scorers as well as test-retest reliability studies. During the third phase, conducted from August through October 2019, we validated the norms between performance of the sample and national norms, calculated the reliability between human and machine scoring, evaluated the passages for differential item functioning (DIF), and conducted validation studies with the Istation's Indicators of Progress Early Reading (ISIP™ ER) assessment and DIBELS 8.

### **Phase One: Training the Acoustic Models**

During the first two phases, Istation recruited schools within our extensive customer base to participate in pilot testing of the ISIP ORF. During Phase One, over 3,000 students in grades 1 through 3 in the Inland South, West, Lowland South, and Midland recorded over 14,000 passages. Most of the students (67.4%) recorded three or more passages. Students had a wide variety of backgrounds, including students who were non-Hispanic white, African-American or black, Hispanic, and other race ethnicities. Students in first grade recorded 4,596 passages, students in second grade recorded 6,081 passages, and students in third grade recorded 4,040 passages.

Istation provided headsets that had noise-cancelling microphones for each participating classroom. Istation requested that students each record three passages in a session and that they participate in at least three sessions. The software presented the passages randomly, and students participating in multiple sessions did not read the same passage twice.

### Phase Two: Passage Difficulty and Reliability of Scoring

After the acoustic models were trained, we then turned to the next phase of assessing item difficulty and reliability. During this phase, over 2,000 students participated in the field-testing, and they made over 17,000 recordings. Over 95% of the students recorded three or more passages with a mean of nine passages per student.

### **Phase Two Sample Characteristics**

Sample characteristics are reported at the school level, based on the number of passage recordings, and these characteristics are reported in Table 3. For socioeconomic status (SES), we used the 2018 criteria for poverty at the school from the National Center for Education Statistics. High poverty schools have greater than 75% participation in free or reduced-price lunch (FRPL). Mid-high poverty schools consist of those with 50.1% to 75% FRPL, mid-low poverty schools are 25% to 50% FRPL, and schools with less than 25% FRPL are categorized as low poverty schools.

To determine the racial/ethnic composition, we created categories based on the percentage of African American, Hispanic, or white students. Schools that served greater than 40% African American or black population percentages were considered to have a large African American student body. We constructed a similar category for students of Hispanic origin. We

created categories based on quartiles for the white student population categories, consisting of 82% white, 55 to 82% white, 15 to 55% white, and less than 15% white.

There was representation across all categories for SES. The sample was slightly over-represented for mid-high poverty schools and under-represented in low-poverty schools. Forty seven percent (47%) of the sample came from schools with greater than 40% Hispanic origin students, and 8.8% came from schools with greater than 40% African American students. There was representation across all categories for percentage of the study body that are white, with some under-representation in schools that are between 55 and 82% white. Overall, we deemed that the sample had adequate representation across the demographic variables.

Table 3. Istation Oral Reading Fluency Demographics – Phase Two

<i>N = 2,006</i>	<i>ORF Phase Two Student Sample</i>	<i>NCES 2016-2017</i>
<b>SES at the school</b>		
<i>High Poverty</i>	24.40%	29.4%
<i>Mid-High Poverty</i>	30.1%	24.4%
<i>Mid-Low Poverty</i>	35.5%	22.6%
<i>Low Poverty</i>	< 1.0%	17.6%
<i>Missing from NCES</i>	10.0%	
<b>Race/Ethnicity at the School</b>		
<i>Greater than 40% African American</i>	8.8%	12.5%
<i>Greater than 40% Hispanic Origin</i>	47.0%	27.9%
<i>Less than 15% White</i>	25.8%	28.9%
<i>15 to 54.9% White</i>	24.5%	27.5%
<i>55 to 82% White</i>	39.1%	24.8%
<i>Greater than 82% White</i>	9.2%	18.7%

### Evaluation of the Passages

In this phase, if students participated in Phase One, they continued to read randomized passages that they had not read before. Teachers or trained scorers listened to the passages and

marked whether the recording was of sufficient quality. All passages were manually scored at least once by teachers, former teachers, or trained scorers, and over 3,100 passages were manually scored twice.

We calculated the WCPM for the passage by taking the number of words the student read correctly, dividing it by the amount of time (in seconds) that the student spoke, and dividing that by 60. If a student got 48 words correct in 24 seconds, then the words correct per minute is 120 (e.g.,  $48/(24/60) = 120$ ). We calculated WCPM for both the human and the automated scoring results.

### Passage Difficulty

We evaluated the passages to determine if different passages would result in higher or lower WCPM scores. Passages were randomized throughout the months of data collection. First, we ran correlational analysis by grade with Lexile ranges and words correct per minute to determine if passage difficulty and WCPM were related. After eliminating passages with less than five WCPM, the correlation was .009 for grade 1 ( $p > .05$ ),  $-.037$  for grade 2 ( $p < .05$ ), and .023 for grade 3 ( $p > .05$ ), indicating that while there was a significant correlation in grade 2, it was not in the expected direction, and the impact was minimal. Therefore, we concluded that there was not a meaningful relationship between passage reading level and WCPM.

Next, we evaluated the mean number of WCPM between passages. In the final version, a standard administration will consist of three passages. The lowest scoring passage is dropped, and the other two will be averaged together to calculate a final score. This method was decided upon because there are instances where the automated scoring will not correspond to a teacher's score due to several possible reasons: background noise in the recording, students' speaking very softly, or other reasons. This method also helps to reduce variability in scores due to the

characteristics of a particular passage or practice effects. We evaluated item pairs that would either inflate or deflate a student's score. Since the score is calculated by averaging the two highest passages and dropping the lowest, we were concerned with passages that would systematically over- or under-inflate a score when combined with other passages. We identified one passage in first grade, one in second grade, and two in third grade that over-inflated scores by more than a standard deviation when paired with any other passage, and we therefore eliminated these items from the pool of passages. In addition, there were three passages in third grade that when paired together deflated scores, and the system was programmed to not pair these passages together.

## **Reliability**

We conducted three types of reliability studies for ISIP ORF: inter-rater reliability between human scoring and automated scoring by region, inter-rater reliability between two human scores, and test-retest reliability. For human-automated and human-human scoring reliability, we conducted correlations at the passage level.

### **Human-Automated Scoring Passage Level**

We compared the WCPM between the human and automated scoring overall and by dialectical region where there were sufficient numbers of passages recorded. We selected scorable passages where there were more than five words recorded. Examination of the data revealed that there were some outliers with high human scores and low automated scores. We randomly selected passages that were more than two standard deviations below the mean and listened to the files. Most had extensive background noise or were silent, indicating data entry



errors. We identified extreme outliers in a simple regression analysis and eliminated those where the studentized residual was less than  $-6.0$ .

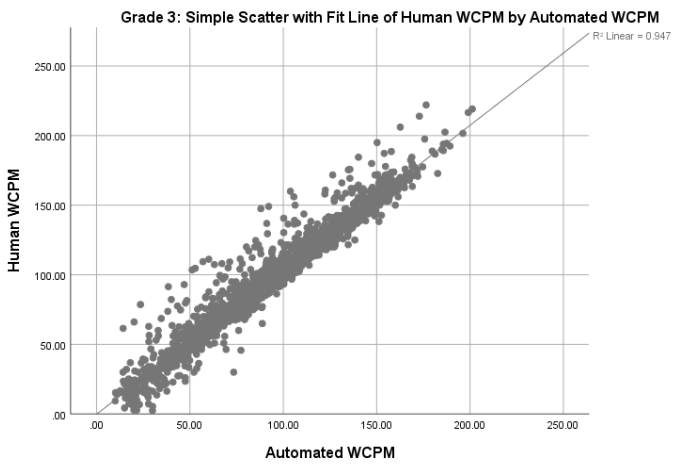
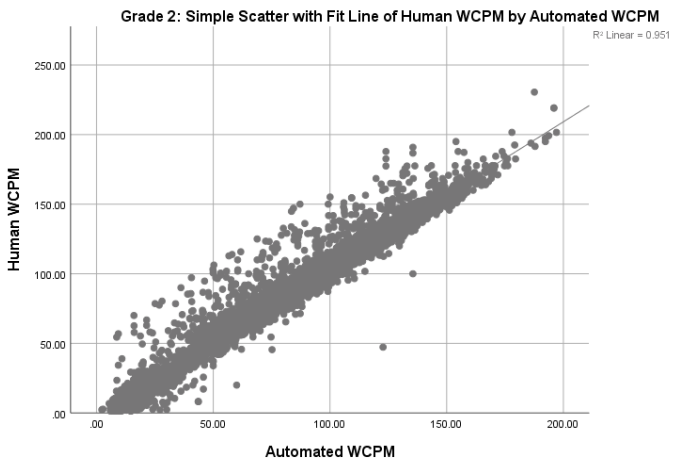
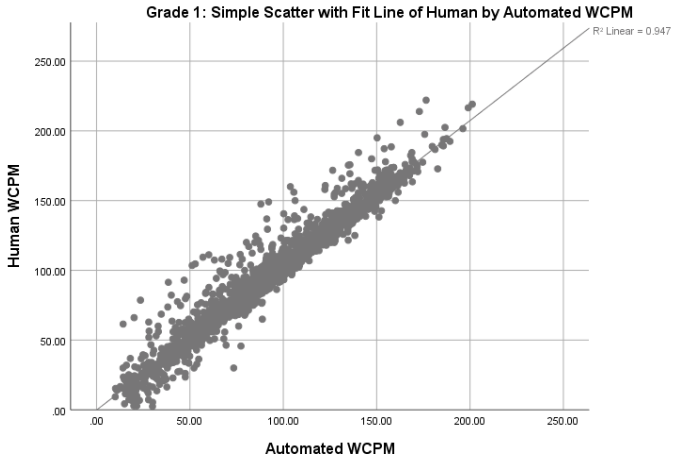
Correlations between human and automated scoring are available in Table 4. Correlations were  $.98$  ( $p < .001$ ) for grade 1,  $.98$  ( $p < .001$ ) for grade 2, and  $.97$  ( $p < .001$ ) for grade 3, all indicating a strong correlation between human and automated scoring using the scoring method described above. There were also strong correlations between human and automated scoring by dialect region: Inland South ( $.95$ ,  $p < .05$ ), Lowland South ( $.95$ ,  $p < .05$ ), and West ( $.99$ ,  $p < .05$ ).

Table 4. ISIP ORF Passages: Correlations for Human and Automated Scoring by Grade and Dialectical Region

	<i>Human vs. Automated Scoring</i>	<i>Inter-Rater Human Scoring</i>	<i>Difference Human-Human Scoring</i>	<i>Difference Human-Automated Scoring (both graders)</i>
<i>Grade 1</i>	$.98^{**}$ (N = 6,527)	$.99^{**}$ (N = 1,697)	$.21$ (N = 1,697)	$.06$
<i>Grade 2</i>	$.98^{**}$ (N = 7,656)	$.99^{**}$ (N = 1,054)	$.09$ (N = 1,054)	$3.7$
<i>Grade 3</i>	$.97^{**}$ (N = 2,756)	$.98^{**}$ (N = 434)	$1.19$ (N = 434)	$4.06$
<i>Inland South</i>	$.95^{**}$ (N = 2,756)			
<i>Lowland South</i>	$.96^{**}$ (N = 4,151)			
<i>West</i>	$.99^{**}$ (N = 5,399)			

$** p < .01$

Figure 2 shows a scatterplot between human and automated scoring by grade. These scatterplots demonstrate a high correlation between the two types of scores.



*Figure 2. Human and Automated Scoring by Grade, March – May Recorded Passages*

## Human-Automated Scoring ISIP ORF Scores

After determining a strong correlation between human and automated scoring at the passage level, we next calculated the ISIP ORF score for each participating student for passages recorded in March, April, and May of 2019. Only students who had recorded three passages in a session were included in this analysis. We evaluated these by month so that we could determine if the students were growing in their reading ability across the spring semester. We expected that younger students would grow more than older students (Hasbrouck & Tindall, 2017), and those expectations have been confirmed. Students in grade 1 had a gain of 8 WCPM from March to May, while students in grade 2 had a gain of just under 3 WCPM, and students in grade 3 had a mean gain of 3 between March and April and then showed a loss in May (note that not all students participated in all three months). After calculating the ISIP ORF score for both the human and automated scores, the correlations remained strong at .90 or higher for all grades, by month of administration.

We also calculated the difference between human and automated scoring, and the effect size is calculated using the standard deviation of the human score. The smallest mean discrepancies were in first grade, and the largest mean discrepancies were in third grade. Effect sizes are small for each grade and month, ranging from .001 to .178. These results are available in Table 5.

Table 5. ISIP ORF Scores: Correlations between Human and Automated Scoring

	<i>Students</i>	<i>Correlation – Human and Automated WCPM Scores</i>	<i>Means and Standard Deviation Automated Scoring</i>	<i>Means and Standard Deviation Human Scoring</i>	<i>Human- Automated Scoring</i>	<i>Effect Size</i>
<i>Grade 1</i>	N = 466	March: .96***	48.06 (27.82)	48.10 (31.22)	.04	.001
	N = 299	April: .99***	53.97 (30.36)	54.02 (33.37)	.05	.001
	N = 226	May: .91***	56.16 (29.69)	58.63 (33.63)	2.47	.073
<i>Grade 2</i>	N = 656	March: .98 ***	74.50 (32.07)	77.29 (34.84)	2.79	.080
	N = 382	April: .96 ***	76.14 (31.27)	79.38 (34.37)	3.24	.094
	N = 206	May: .95 ***	77.38 (32.99)	80.38 (35.68)	3.00	.084
<i>Grade 3</i>	N = 266	March: .97***	90.11 (33.58)	92.99 (34.86)	2.88	.083
	N = 180	April: .97***	93.39 (31.04)	96.78 (33.00)	3.39	.103
	N = 159	May: .90***	90.78 (31.99)	96.75 (33.50)	5.97	.178

\*\*\*  $p < .001$

### Test-Retest Reliability

Students who participated in the ISIP ORF assessment in adjacent calendar months of March and April were included in a test-retest reliability study. After we calculated the ISIP ORF score, we evaluated the mean scores, which increased by each grade and each month, with greater growth in first and second grades. Pearson correlations were run for test-retest reliability, and the correlations between scores were strongly correlated. Results are available in Table 6.

Table 6. Test-Retest Reliability and Means, Standard Deviations of Automated WCPM

<i>Grade</i>		<i>MAR</i>	<i>APR</i>	<i>Reliability</i>
1	Mean	49.43	52.59	.92***
	<i>SD</i>	26.88	28.01	
	<i>N</i>	213	213	
2	Mean	74.56	76.99	.90***
	<i>SD</i>	31.49	30.89	
	<i>N</i>	267	267	
3	Mean	92.00	93.30	.85***
	<i>SD</i>	30.87	30.76	
	<i>N</i>	101	101	

\*\*\* $p < .001$

### Summary of Phase Two Results

The ISIP ORF has passages that produce consistent results, and there are strong correlations between human and automated scoring. When there are differences between the two scoring methods, the human raters produce scores that are slightly higher than the automated scoring, but the effect sizes are small. There is also strong reliability for test-retest, evidenced by strong correlations for first, second, and third grade; and consistent with classroom expectations, there is evidence of reading growth across the two months.

## Phase Three – Fall 2019

Given the reliability of results in the ISIP ORF pilot, Istation released the ISIP ORF for use by customers in the fall of 2019 for manual scoring. Automated scoring was implemented in a region after it met the criteria of having at least 10,000 voice samples, and having reliability between human and automated scoring of .95 and greater. Automated scoring was activated in the West, Inland South, and Lowland South in October 2019.

During Phase Three, over 75,747 students in grades 2 and 3 recorded 298,220 passages, of which 70,687 were scored by teachers. These recordings were predominately in the Inland South and Lowland South regions. Istation recommended the use of noise-cancelling headphones, but using a standard laptop microphone was permissible during this phase.

Phase Three had four primary research goals. First we wanted to evaluate whether the ISIP ORF resembled national norms composed by Hasbrouck and Tindall (2017), the most commonly used norms. Second, we wanted to evaluate if there was differential item functioning by race/ethnicity and gender. Third, we wanted to evaluate the correlation between human and automated scoring with the relaxed criteria regarding headphone use. And fourth, we evaluated concurrent validity with the ISIP ER and the DIBELS 8.

### **Phase Three Sample Characteristics**

To evaluate our research goals, we first selected passages that were recorded between mid-August and early October 2019. Our requirements were that a student had three passages from which we computed the ISIP ORF score by dropping the lowest passage and averaging the other two. Next, we randomly selected students to represent a national sample based on individual level race/ethnicity and gender, and socioeconomic status at the school level, using the

random selection package in *R* statistical software. Race/ethnicity had four categories: Non-Hispanic white, Non-Hispanic African-American or black, Hispanics of any race, and Asian and other races/ethnicities. Socioeconomic status was divided into four levels using the NCES criteria for high, mid-high, mid-low, and low-poverty schools described in Phase Two. We obtained sample targets from the NCES 2016-2017 school year, the latest available. The norming sample consisted of 10,000 students in grades 2 and 3 (5,000 in grade 2 and 5,000 in grade 3), and the demographics are available in Table 7. All demographics closely or exactly match the national target.

### **DIF Analysis**

To determine whether there was a racial/ethnic or gender bias in the items, we conducted differential item functioning analysis, or DIF, which is described as the difference in an item's difficulty between subgroups of examinees who have the same ability level on the trait being measured (Patarapichayatham, Kamata, & Kanjanawasee, 2011). DIF occurs when an item on a test functions differently for different groups, given the same ability level. Each passage was treated as an item. The logistic regression DIF detection method is applied to detect uniform DIF, which is divided into three levels: Level A items have negligible or non-significant DIF effects, B items have moderate DIF, and C items have large DIF effects. We calculated two DIF factors: gender (male vs. female) and race (Non-Hispanic white vs. all other combined). The analysis used logistic regression DIF detection analysis by *difR* package in *R*. Total WCPM scores were used as the matching criteria. *DifR* obtained two DIF detection criteria: Zumbo & Thomas (ZT) and Jodoin & Gierl (JG). Both criteria had the same procedure but different cut points. We conducted the analysis with the norming sample. Using Zumbo & Thomas (1990)

Table 7. Demographic Characteristics of the Phase Three Sample

<b>Gender</b>		
<i>Male</i>	Grade 2 ISIP ORF	50.0%
	Grade 3 ISIP ORF	50.4%
<i>Female</i>	Grade 2 ISIP ORF	50.0%
	Grade 3 ISIP ORF	49.6%
<b>Race/Ethnicity</b>		
<i>White, Not Hispanic</i>	National %	46.0%
	Grade 2 ISIP ORF	46.0%
	Grade 3 ISIP ORF	46.0%
<i>African American, Not Hispanic</i>	National %	15.6%
	Grade 2 ISIP ORF	15.6%
	Grade 3 ISIP ORF	15.6%
<i>Hispanic</i>	National %	27.8%
	Grade 2 ISIP ORF	27.8%
	Grade 3 ISIP ORF	27.8%
<i>Asian/Other</i>	National %	11.6%
	Grade 2 ISIP ORF	11.6%
	Grade 3 ISIP ORF	11.6%
<b>Socioeconomic Status</b>		
<i>SES 1</i> Greater than 75% FRPL	National %	35.3%
	Grade 2 ISIP ORF	35.3%
	Grade 3 ISIP ORF	35.3%
<i>SES 2</i> 50.1-75% FRPL	National %	24.4%
	Grade 2 ISIP ORF	24.4%
	Grade 3 ISIP ORF	24.4%
<i>SES 3</i> 25-50% FRPL	National %	22.6%
	Grade 2 ISIP ORF	22.6%
	Grade 3 ISIP ORF	22.6%
<i>SES 4</i> Less than 25% FRPL	National %	17.7%
	Grade 2 ISIP ORF	17.7%
	Grade 3 ISIP ORF	17.7%

*Target percentages derived from the National Center for Education Statistics, 2016-2017*

*Common Core data file.*



DIF criterion, results showed that all of the passages displayed as A items with negligible or non-significant DIF effects. There were no B or C items. Overall, our ORF passages did not show any potential bias.

### **Automated Scoring and Human Scoring**

Not all passages were evaluated by human scorers during this phase; therefore, we listened to a sample of passages to determine if the quality would be reliable for an automated score. After listening to sample passages, we noticed the majority of recordings were of good quality; however, we noted the following characteristics that may be problematic when accepting a final score:

- students' speaking softly or whispering that FLORA did not recognize
- noisy backgrounds with several voices audible
- students with heavy accents that FLORA did not recognize
- teachers' giving instructions to the student while the student was trying to read

To help identify these possibly out-of-pattern recordings, we set the criteria at 30 words spoken for a high confidence passage, and less than 30 words will require review to ensure that the automated scoring is accurate. Using all available scores from the fall of 2019, and the ISIP ORF scoring criteria, correlations are moderate for review scores and strong for high confidence scores, indicating that when the 30-words-spoken criteria is met, the acoustic models are reliable. These results are available in Table 8.

Table 8. Correlations of ORF Score with Human Score: Fall Data

	<i>High Confidence</i>	<i>Review Scores</i>
<i>Grade 2</i>	0.90** (N = 6,438)	0.66** (N = 4,110)
<i>Grade 3</i>	0.89** (N = 8,450)	0.65** (N = 2,728)

\*\* $p < .01$

The high confidence correlations, while strong, are lower than those achieved in the second research phase. During that phase, administration was standardized with the use of noise-cancelling headphones, and a human listened to each passage and recorded whether the passage was scorable. These correlations are similar to the research conducted by Bolaños et al. (2013).

The results from both phases indicate that there is a strong correlation between human and automated scoring. Quieter environments yield more reliable recordings. We did not conduct extensive data cleaning in this analysis because when the ISIP ORF is used in schools to assess students, it is important to include all types of recordings, including those with noisy backgrounds or otherwise inadequate recording environments. There were some cases in our data that had automated scores without teacher scores above 0 (null) when there was manual scoring. While some of these may be data entry errors, it is possible that FLORA will recognize sounds or words in a noisy environment that match passages and record a score. Using all available cases, we evaluated the difference between the high confidence and review scores, and the majority of these types of recordings are in the review score category (Chi Square = 3,784.1,  $p < .001$ ).

Results are available in Table 9.

Table 9. Automated Scoring for Review and High Confidence Scores and Null Teacher Scoring

		<i>Teacher Score &gt; 0</i> N = 19,666	<i>Teacher Score = 0</i> N = 2,060
<i>Review Scores</i>	N = 6,838	72.5%	27.5%
<i>High Confidence Scores</i>	N = 14,888	98.8%	1.2%

---

*Chi Square = 3,784.1, p < .001*

While most of the review scores are reliable with correlations between .66 and .65, the majority of the null teacher scores are in this category. With high confidence scores, slightly more than 1% would not earn a teacher score due to the recording quality. These results point to the importance of having a more controlled environment when students are recording their passages. Istation highly recommends the use of noise-cancelling headphones and computer labs to help maintain a more conducive recording environment. Out-of-pattern scores, either with a score that is much higher or much lower than a teacher expects a student would achieve, should be manually reviewed with the ISIP ORF system to verify that the score is representative of the student’s true oral reading ability.

### **Evaluation of the National Norms**

Hasbrouck and Tindal (1992, 2006, 2017) developed national norms for oral reading fluency in 1992 and 2006 and updated them in 2017. Until that time, students and districts composed local norms. The 1992 norms were compiled for grades 2-5 for fall, winter, and spring, and they were computed for the 75th, 50th, and 25th percentiles. In 2006 they were updated to include first grade starting in winter, and the sample was based on students in 23

states. These national norms became the standard across the country and were published to prevent a low-performing school from setting cut points that were too low.

In 2017 Hasbrouck and Tindal updated the norms using information from standardized assessments including DIBELS 6th Edition©, DIBELS NEXT©, and easyCBM© ORF assessments. Hasbrouck and Tindal compiled the data from these assessments into a master file and averaged the scores with a grade, season, and assessment. The authors did not report each percentile rank; rather they reported the percentile cut points at the 90th, 75th, 50th, 25th, and 10th percentiles, similar to the previous norms. The 2017 norms are now the most up-to-date standard for oral reading fluency. These compiled norms allow educators to interpret a student's ORF performance outside of a specific assessment product.

We evaluated whether the average scores for the scores collected in Phase Three corresponded to the national norms created by Hasbrouck and Tindal (2017b). Using the Phase Three sample, we calculated percentile ranks using the automated WCPM scores and compared them to national norms. Results are available in Table 10. The ISIP ORF percentiles correspond closely to the Hasbrouck and Tindal (2017) norms at the critical cut points in the 10th, 25th, and 50th percentiles, indicating that the ISIP ORF yields comparable results as other ORF assessments. These results confirm research that ORF is a stable and robust measure that is remarkably durable (Hasbrouck & Tindal, 2017b), and given adequate conditions for testing and recording quality, users of ISIP ORF can use the Hasbrouck and Tindal (2017b) national norms for first, second, and third grades. This is advantageous to users, as it will allow them to compare performance of their students to the national standard.

Table 10. National Norms and Fall 2019 Percentiles from ISIP ORF

		<i>Hasbrouck &amp; Tindal 2017 Percentiles</i>	<i>ISIP ORF Fall 2019 Percentiles</i>
<i>Grade 2</i>	90 <sup>th</sup> percentile	111.0	109.0
	75 <sup>th</sup> percentile	84.0	84.0
	50 <sup>th</sup> percentile	50.0	57.5
	25 <sup>th</sup> percentile	36.0	35.0
	10 <sup>th</sup> percentile	23.0	23.0
<i>Grade 3</i>	90 <sup>th</sup> percentile	134.0	134.0
	75 <sup>th</sup> percentile	104.0	107.5
	50 <sup>th</sup> percentile	83.0	85.0
	25 <sup>th</sup> percentile	59.0	60.0
	10 <sup>th</sup> percentile	40.0	39.0

### **Concurrent Validity**

The standards for educational testing state that the primary source of validity evidence is the extent to which the relationships implied by the theoretical construct are confirmed (American Educational Research Association, 2014). We conducted correlational analysis between the ISIP ORF and the ISIP ER and DIBELS 8. The evidence with ISIP ER establishes that ISIP ORF is a measure of reading ability that corresponds with other areas of the Big Five, and the evidence with the DIBELS 8 establishes that ISIP ORF has concurrent validity with other ORF measures.

## Validity with ISIP ER

ORF is a measure of oral reading that is correlated with other reading skills such as vocabulary and comprehension. Fluent readers have more automaticity, allowing them to read faster with comprehension, and greater vocabulary helps them read more efficiently. We expect that ISIP ORF will correlate with other reading measures, including the ISIP ER. The vast majority of students in the Phase Three sample took an ISIP ER assessment in August, September, or October. The ISIP ER is a computer-adaptive test that delivers continuous progress monitoring for students in early elementary school. We computed correlations with the overall ISIP score as well as the Reading Comprehension, Text Fluency, Spelling, and Vocabulary subtests. In Reading Comprehension, students match sentences and pictures and complete sentences by choosing the best word to complete the text. Text Fluency measures students' ability to read with comprehension, and items use a maze task where students are asked to select words that make the most sense in the sentence. Spelling is assessed by giving students an array of letters and asking them to spell a specific word using the letters. Vocabulary is designed to test students' knowledge of word meanings that are not typically used in conversation but are frequently encountered while reading (Mathes, Torgeson, & Herron, 2016).

For the ISIP ER, we computed correlations using the Phase Three sample, available in Table 11. The correlations were computed by month, so recordings conducted in September were correlated with September ISIP ER scores, and recordings conducted in October were correlated with October ISIP ER scores. We did not compute August recordings with ISIP ER scores as there was insufficient data for the ISIP ER. Overall, students had slightly higher correlations in September than in October, with the exception of grade 3 vocabulary. Students in grade 2 also had slightly higher correlations than students in grade 3 on the overall ISIP as well as the

Comprehension and Fluency subtests. This is consistent with research that shows a stronger relationship between ORF and reading comprehension in earlier grades (Valencia et al., 2010). There were moderate to strong correlations for the overall score, which ranged from .67 to .71, and Reading Comprehension, which ranged from .67 to .72. Correlations were moderate for the Spelling, Text Fluency, and Vocabulary subtests.

Table 11. Correlations with the ISIP ORF and ISIP Overall and Subtest Scores by Grade and Month

	<i>ISIP Overall</i>	<i>ISIP Comprehension</i>	<i>ISIP Text Fluency</i>	<i>ISIP Spelling</i>	<i>ISIP Vocabulary</i>
<b>September</b>					
<i>Grade 2 ISIP ORF</i>	.71** (N = 2,948)	.72** (N = 2,947)	.70** (N = 2,948)	.62** (N = 2,946)	.52** (N = 2,947)
<i>Grade 3 ISIP ORF</i>	.68** (N = 2,916)	.68** (N = 2,916)	.62** (N = 2,916)	.62** (N = 2,918)	.47** (N = 2,918)
<b>October</b>					
<i>Grade 2 ISIP ORF</i>	.68** (N = 418)	.69** (N = 423)	.68** (N = 418)	.57** (N = 423)	.49** (N = 429)
<i>Grade 3 ISIP ORF</i>	.67** (N = 451)	.67** (N = 451)	.67** (N = 451)	.59** (N = 453)	.50** (N = 456)

\*\* $p < .01$

### Validity with DIBELS 8

We also evaluated concurrent validity with DIBELS 8. There were 38 students in second grade and 49 students in third grade. Within each grade, half of students took the DIBELS ORF fall benchmark assessment first and then ISIP ORF. The other half of students took the ISIP ORF assessment first and then DIBELS. The data were collected in November and December 2019 in

the states of Texas, New Mexico, New York, and Virginia. The sample consisted of 50.6% males and 49.4% females. The sample was 47.1% Non-Hispanic whites, 11.5% African-American or black, 32.2% of Hispanic origin, and the remainder were Asian or other race/ethnicities. The concurrent validity coefficient between ISIP ORF accuracy and DIBELS ORF accuracy was 0.83, indicating a strong relationship between ISIP ORF and DIBELS ORF. An accuracy is computed from this formula:  $Accuracy = \left[ \frac{wordscorrect}{wordscorrect + wordserrors} \right] * 100$ . Results are available in Table 12. The concurrent validity coefficient between ISIP ORF WCPM and DIBELS ORF WCPM was 0.89, indicating a very strong relationship between ISIP ORF and DIBELS ORF.

Table 12. Correlations between ISIP ORF and DIBELS 8, WCPM and Accuracy

	<i>DIBELS 8 WCPM</i>	<i>DIBELS 8 Accuracy</i>
<i>ISIP ORF</i>	.89**	.83**

\*\*  $p < .01$

### **Guidelines for Administering ISIP ORF**

Oral reading fluency is an important component of the Big Five, and the science of reading has established its durability for use in the classroom as an indicator of reading progress (Hasbrouck & Tindal, 2017a). Given ORF’s brevity and reliability, they are appropriate for use as a progress-monitoring tool and for universal screening of reading difficulties. The ISIP ORF is easy to administer and engaging for students as the assessment is designed to simulate a



recording studio with the student as voice talent. Passages are grade leveled and offer a mixture of fiction and nonfiction, and there are sufficient passages for students to be assessed in nine separate sessions without reading the same passage twice.

The ISIP ORF can be used to assess student progress according to the national norms provided by Hasbrouck and Tindal (2017b). Automated scoring is currently available in the Inland and Lowland South and West dialect regions for grades 2 and 3. Recording capabilities, passages, and manual scoring are available for all regions and for students in kindergarten and first grade.

Automated scoring for ISIP ORF is reliable and valid, and there are recording conditions identified to enhance its effectiveness. Throughout our field tests, there were instances of students recording passages in noisy classrooms or labs, and these environments impaired the ability of the acoustic models to accurately score the passages. When headsets with noise-cancelling microphones were used, these problems were reduced and the recordings were of much higher quality. While a computer microphone can be used effectively if the surrounding environment is quiet, Istation recommends using headsets.

While the acoustic models have high accuracy, especially in the high confidence passages where more than 30 words were recorded, there were instances in the data where the automated scoring was significantly higher than manual scoring. These results were due in large part to noisy backgrounds. Teachers should review the results from the ISIP ORF carefully when using automated scoring. Based on their expert judgment and knowledge of the student, teachers are advised to listen to recordings that do not appear to accurately reflect students' reading ability.

## **Future Development**

Istation has continued development goals for ISIP ORF. Currently we have passages and manual scoring available for kindergarten and grade 1, and passages for grades 4 and 5 are forthcoming. Automated scoring for grade 1 and additional dialect regions will be released after there are sufficient passages recorded to train the acoustic models to a level of acceptable accuracy. Future development will include passages and manual scoring in Spanish for students in kindergarten through fifth grade.

## REFERENCES

- American Educational Research Association, American Psychological Association & National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, D. C.: AERA.
- Aschmann, R. (2018). North American English Dialects, Based on Pronunciation Patterns. [www.aschmann.net/AmEng/](http://www.aschmann.net/AmEng/). Retrieved 12/13/2019.
- Bolaños, D., Cole, R. A., Ward, W. H., Tindal, G. A., Hasbrouck, J., & Schwanenflugel, P. J. (2013). Human and Automated Assessment of Oral Reading Fluency. *Journal of Educational Psychology*, *105*(4), 1142–1151. <https://doi.org/10.1037/a0031479>
- Eason, S. H., Sabatini, J., Goldberg, L., Bruce, K., & Cutting, L. E. (2013). The utility and accuracy of oral reading fluency score types in predicting reading comprehension: Examining the relationship between word reading efficiency and oral reading rate in predicting comprehension among different types of readers. *Scientific Studies of Reading*, *17*(3), 199–223.
- Hasbrouck, J., & Tindal, G. (1992). Curriculum-based oral reading fluency norms for students in grades 2-5. *Teaching Exceptional Children*, *24*(3), 41-44.
- Hasbrouck, J., & Tindal, G. (2006). Oral Reading Fluency Norms: A Valuable Assessment Tool for Reading Teachers. *The Reading Teacher*, *59*(7), 636-644. Retrieved from [www.jstor.org/stable/20204400](http://www.jstor.org/stable/20204400)
- Hasbrouck, J. & Tindal, G. A. (2017a). Oral Reading Fluency: Outcomes from 30 Years of Research (Technical Report No. 1701). Behavioral Research and Teaching, University of Oregon. Retrieved from <http://brt.uoregon.edu>
- Hasbrouck, J. & Tindal, G. (2017b). *An update to compiled ORF norms* (Technical Report No. 1702). Eugene, OR, Behavioral Research and Teaching, University of Oregon. Retrieved from <http://brt.uoregon.edu>
- Hintze, J. M., Callahan III, J. E., Matthews, W. J., S. Williams, S. A., & Tobin, K. G. (2002). Oral Reading Fluency and Prediction of Reading Comprehension in African American and Caucasian Elementary School Children. *School Psychology Review*, *31*(4), 540. Retrieved from a9h.
- Labov, W., Ash, S., & Boberg, C. (2006). *The Atlas of North American English: Phonetics Phonology and Sound Change. A Multimedia Reference Tool*. Berlin: Walter de Gruyter GmbH & Co.
- Magis, D., Beland, S., & Raiche, G. (2018). *difR package: Collection of Methods to Detect Dichotomous Differential Item Functioning (DIF)*.

- R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Mathes, P., Torgesen, J., & Herron, J. (2016). Istation's Indicators of Progress (ISIP) Early Reading: Technical Report. Dallas, TX: Istation.
- National Center for Education Statistics (2017). Common Core of Data. Washington, D.C.: National Center for Education Statistics. Retrieved from [nces.ed.gov/ccd](https://nces.ed.gov/ccd).
- National Reading Council, (1998). Preventing Reading Difficulties in Young Children. Washington, D. C.: National Academy of Sciences. Retrieved from [files.eric.ed.gov](https://files.eric.ed.gov).
- National Reading Panel, (2000). *Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and its Implications for Reading Instruction*. Washington, D.C.: National Institutes of Health.
- Patarapichayatham, C., Kamata, A. & Kanjanawasee, S. (2012). Evaluation of Model Selection Strategies for Cross-Level Two-Way Differential Item Functioning Analysis. *Educational and Psychological Measurement, 72 (1) 44-51*. doi/10.1177/0013164411409743.
- Pearce, L. R., & Gayle, R. (2009). Oral Reading Fluency as a Predictor of Reading Comprehension With American Indian and White Elementary Students. *School Psychology Review, 38(3)*, 419–427. Retrieved from a9h.
- Petscher, Y., & Kim, Y.-S. (2011). The utility and accuracy of oral reading fluency score types in predicting reading comprehension. *Journal of School Psychology, 49(1)*.
- Quirk, M., & Beem, S. (2012). Examining the relations between reading fluency and reading comprehension for English language learners. *Psychology in the Schools, 49(6)*, 539–553. Retrieved from a9h.
- Schilling, S. G., Carlisle, J. F., Scott, S. E., & Zeng, J. (2007). Are Fluency Measures Accurate Predictors of Reading Achievement? *The Elementary School Journal, 107(5)*, 429–448. <https://doi.org/10.1086/518622>
- Valencia, S. W., Smith, A. T., Reece, A. M., Li, M., Wixson, K. K., & Newman, H. (2010). Oral Reading Fluency Assessment: Issues of Construct, Criterion, and Consequential Validity. *Reading Research Quarterly, 45(3)*, 270–291. Retrieved from JSTOR.
- Wanzek, J., Al Otaiba, S., & Petscher, Y. (2014). Oral Reading Fluency Development for Children With Emotional Disturbance or Learning Disabilities. *Exceptional Children, 80(2)*, 187-204. <https://doi.org/10.1177/001440291408000204>. Retrieved from a9h.