Published in final edited form as:
Read Psychol. 2012 ; 33(1-2): 133–161. doi:10.1080/02702711.2012.631863.

Reliability and Validity of Oral Reading Fluency Median and Mean Scores among Middle Grade Readers When Using Equated Texts

Amy E. Barth,
Department of Psychology, Texas Institute for Measurement, Evaluation and Statistics, and the Texas Center for Learning Disabilities, University of Houston

Karla K. Stuebing,
Department of Psychology, Texas Institute for Measurement, Evaluation and Statistics, and the Texas Center for Learning Disabilities, University of Houston

Jack M. Fletcher,
Department of Psychology and the Texas Center for Learning Disabilities, University of Houston

Paul T. Cirino,
Department of Psychology, Texas Institute for Measurement, Evaluation and Statistics, and the Texas Center for Learning Disabilities, University of Houston

Melissa Romain,
Department of Psychology, Texas Institute for Measurement, Evaluation and Statistics, and the Texas Center for Learning Disabilities, University of Houston

David Francis, and
Department of Psychology, Texas Institute for Measurement, Evaluation and Statistics, and the Texas Center for Learning Disabilities, University of Houston

Sharon Vaughn
Meadows Center for Preventing Educational Risk, University of Texas-Austin, Austin TX

Abstract

We evaluated the reliability and validity of two oral reading fluency scores for one-minute equated passages: median score and mean score. These scores were calculated from measures of reading fluency administered up to five times over the school year to students in grades 6–8 (n = 1,317). Both scores were highly reliable with strong convergent validity for adequately developing and struggling middle grade readers. These results support the use of either the median or mean score for oral reading fluency assessments for middle grade readers.

Keywords

response to intervention; reading fluency; middle grade readers

Oral Reading Fluency

Oral reading fluency (ORF), a measure of accuracy and rate of reading grade level text, is a component of progress monitoring growth in reading during early elementary school (grades

Correspondence concerning this article should be addressed to Amy E. Barth, University of Houston, Department of Psychology, 2151 West Holcombe, 222 Texas Medical Center Annex, Houston, Texas 77204-5053. amy.barth@times.uh.edu.

1–3). In general, students are individually administered a passage (expository or narrative) and are asked to read aloud for one minute while teachers record errors and determine the number of correct words read during the allocated time (Reschly, Busch, Betts, Deno, & Dong, 2009; Shinn, 1989; Stecker, Fuchs, & Fuchs, 2005; Wayman, Wallace, Wiley, Ticha, & Espin, 2007). This score is plotted in relation to benchmarks to indicate reading progress. The reading materials used for ORF are derived from grade level or instructional level texts selected from the classroom curriculum, basal readers, or pre-packaged ORF texts.

Technical Adequacy of ORF among Elementary Grade Readers

Elementary grade teachers have widely adopted ORF as a primary means of measuring growth in reading. ORF measures have good reliability (Deno, 1992; Reschly, Busch, Betts, Deno, & Long, 2009). This is evidenced from controlled studies that show when teachers use ORF assessment in a curriculum-based measurement (CBM) framework for instructional planning and adaptation, goal setting, and other aspects of instruction, student outcomes are better than when ORF measures are not used (Stecker et al., 2005). In terms of validity, ORF is highly predictive of performance on reading comprehension measures even though it does not directly assess reading comprehension abilities (Hintze & Silberglitt, 2005; McGlichey & Hixson, 2004; Kranxler, Brownell, & Miller, 1998; Shinn, Good, Knutson, Tilly, & Collins, 1992; Stage & Jacobson, 2001). Correlations of ORF and standardized measures of reading comprehension commonly range from 0.50 to 0.90 with most falling around 0.70 for early grade readers (Deno et al., 1982; Marston, 1989; Shinn, 1989). In addition, small changes in reading growth can be reliably quantified to monitor reading progress (Fuchs, Deno, & Mirkin, 1984; Fuchs & Fuchs, 1986; Fuchs, Fuchs, Hamlett, & Ferguson, 1992; Mirkin, Deno, Tindal, & Kuehnle, 1992; Stecker & Fuchs, 2000) and determine whether instruction is beneficial (Powell-Smith & Stewart, 1998; Shinn, Powell-Smith, Good, & Baker, 1997; Tindal, 1989).

Technical Adequacy of ORF among Older Readers

Despite the large research base on the technical adequacy of ORF among early grade readers and its usefulness as a method of informing instruction for elementary grade teachers, gaps in research and practice remain (Ticha, Espin, & Wayman, 2009). One gap is the reliability and validity of ORF among middle grade readers. However, recent research has begun to investigate the technical adequacy of ORF among middle grade readers. For example, Fuchs, Fuchs, and Maxwell (1988) examined the criterion validity of ORF among 70 male students with mild disabilities (age 9–15 years). For the entire sample, the correlation between the average number of words read correct per minute for two passages and scores on the Stanford Achievement Test was 0.91. However, correlations were not reported separately for students in the middle grades.

Building on the findings by Fuchs et al (1988), Espin and Foegen (1996) investigated the relative strength of ORF, maze, and vocabulary matching as a means of predicting middle school students' (n = 184) performance on three researcher developed assessments that tapped comprehension, acquisition, and retention of expository texts. Results indicate that the correlation of ORF and the three reading comprehension tasks ranged from 0.52 to 0.57. Regression analyses indicated that oral reading fluency rates did not uniquely account for variance in reading comprehension after controlling for maze and vocabulary performance.

Yovanoff, Duesbery, Alonzo, and Tindal (2005) investigated the relative importance of vocabulary and oral reading fluency as dimensions of reading comprehension in two independent samples of students (n = 3,203 and n = 3,225) in grades 4–8. ORF was assessed with a one minute read aloud that consisted of a 250 word, grade-level appropriate passages (as determined by the Lexille framework and Flesch-Kincaid readability statistics). Reading

comprehension was assessed with a second grade-level appropriate passage that included 15 comprehension questions. Results revealed that while ORF significantly relates to reading comprehension, its effect diminishes significantly across grades. Correlations between ORF and comprehension ranged from 0.42 to 0.52 in grades 6–8, in contrast to 0.60 to 0.65 in grades 4–5.

Silberglitt, Burns, Madyn, and Lail (2006) evaluated the relationship between ORF and performance on the Minnesota Comprehensive Assessment of Reading among students in 7th (n=582) and 8th (n=843) grade. Students were administered three passages that were standardized and equated using Lexille scores and student performance data (Christ & Silberglitt, 2005). Results indicated that the correlation between the median ORF score and performance on the Minnesota state reading test was 0.60 for students in 7th grade and 0.50 for students in 8th grade.

More recently, Espin, Wallace, Lembke, Campbell, and Long (2010) examined the reliability and predictive validity of ORF among 236 8th grade students. In the fall, students read two passages aloud for three minutes, with the number of words read correctly calculated at one-, two-, and three-minute time frames. All passages were selected from the newspaper, were approximately 800 words in length, and ranged in difficulty from 5th to 7th grade levels, according to Flesh-Kincaid readability formulae. In the winter, students also completed the Minnesota Basic Standards Test (MBST) in reading. Results revealed that the alternate form reliability for the mean ORF score ranged from 0.94 to 0.96 across the one-, two-, and three-minute time frames. Correlations between the mean ORF score and scores on the MBST ranged from 0.78 to 0.79 across the three time frames. For a small subset of students (n=31) ORF growth was examined across ten weeks. Alternate form reliability between adjacent pairs of passages ranged from 0.79 to 0.92 for the mean score obtained at the one minute time frame. Growth over time was minimal for both the one- and three-minute time frames.

In a follow up study, Ticha, Espin, and Wayman (2009) examined the validity and reliability of ORF as an indicator of performance on the Minnesota Basic Skills Test (MBST) and the Woodcock-Johnson III Test of Achievement (WJ-III), Passage Comprehension subtest, among 35 students in 8th grade. Students completed three ORF passages, the WJ-III Passage Comprehension subtest, and the MBST at pretest. For the subsequent 10 weeks, students received one ORF passage. Students read each passage for three minutes with the number of words read correctly calculated at the one-, two-, and three-minute time frames. Passages were created from the local newspaper, were approximately 750 words long, and the readability level as measured by Flesh-Kincaid ranged from fifth to eighth grade levels. Results indicated that the alternate-form reliabilities, as calculated by examining correlations among the scores for the three passages administered at pretest, ranged from 0.95–0.97 for the one-, two-, and three-minute time frames. Differences across time frames were not statistically significant, thus demonstrating relatively little growth over the 10 week progress monitoring period. Correlations between the WJ-III Passage Comprehension and the median ORF score ranged from 0.87 to 0.89 across the one, two, and three minute time frames. Correlations between the MBST and the median ORF score ranged from 0.77 to 0.78. Results indicated that the one minute ORF sample was as reliable and valid as the two or three minute ORF samples.

Limitations of Past Research

Taken together, these studies have begun to address the reliability and validity of ORF measures among middle grade readers. Alternate form reliability ranged from 0.79 to 0.92 (Espin et al., 2009). The correlation between ORF and measures of reading comprehension

(validity) ranged from 0.42 to 0.78, depending on how reading comprehension was measured (Espin et al., 2009; Ticha et al., 2009; Silberglitt et al., 2006; Yovanoff et al., 2005).

However, looking across this body of research, two key features of measures considered to be standardized, longitudinal assessments of reading abilities varied across these studies. The first variation relates to the administration procedures (Fuchs, Fuchs, & Compton, 2004). Note that Silberglitt et al., (2006) and Ticha et al., (2009) used the median score of three passages to calculate the number of words read correct. In contrast, Espin et al., (2009) and Fuchs et al., (1988) used the mean score of two passages to determine students' ORF scores. Thus, it is very likely that the use of different statistics (mean versus median) could lead to different estimates of reliability and validity at the beginning of the school year and across progress monitoring time points.

The second variation is the use of alternate forms that function as equivalent within grade (Jenkins, Fuchs, van den Broek, Espin, & Deno, 2003). Only Silberglit et al., (2006) reported that equating procedures were utilized to ensure that the standard error of measurement was minimal across passages administered fall, winter, and spring (within grade). All other studies utilized readability formulae to determine the equivalence of passages which are problematic because these formulae are imprecise estimates of form equivalence (Francis et al., 2008).

Research on the equivalence of alternate forms has shown that ORF passages deemed equivalent via readability formulae, such as Flesch-Kincaid, vary significantly within grade. For example, Francis et al. (2008) examined the effects of passage order and presentation order on ORF among second grade students (n = 134) randomly assigned to read six passages in one of six fixed orders. Oral reading fluency (WCPM) rates varied across passages but not as a function of passage order. Passage effects (i.e., difficulty level, text type) affected both the shape of ORF growth trajectories and estimates of linear growth rates. However, when the alternate forms were equated, using equipercentile equating methods, the passage effects were removed. Consequently, Francis et al. (2008) suggested that when using ORF to monitor reading progress, equating (e.g., the statistical process of determining comparable ORF scores on alternate forms) is essential to ensure that changes in reading performance from one testing time point to another reflect true change in reading proficiency and do not reflect differences in the difficulty of passages administered. Therefore, to address the issue of form equivalence, this study will used linearly equated scores in order to more precisely estimate the reliability and validity of ORF scores among middle grade readers.

Use of ORF among Struggling Readers

Another issue with any progress monitoring assessment is the extent to which the tool is technically adequate for students with reading disabilities and reading difficulties. Middle school teachers have become increasingly interested in monitoring student's response to general education instruction and special education remedial intervention requiring more precise measurement data (Espin et al., 2010; Ticha et al., 2009). In order to better address the instructional needs of readers who fail to respond to instruction, middle grade teachers are increasingly using ORF to identify at-risk students, inform instructional decision making, and monitor the reading progress of struggling readers and students with disabilities.

Yet little is known about the psychometric properties of ORF among struggling reader populations, particularly at the middle grade level where the correlation between fluency and comprehension is known to be weaker (Jenkins & Jewell, 1993; Yovanoff et al., 2005).

When using ORF we assume that the underlying distribution of fluency is normal. However, using ORF among struggling readers and students with disabilities is likely to be positively skewed because of restriction of range. Restriction of range may occur because design or circumstances (e.g., natural attrition, explicit selection, or incidental selection) abbreviate the values of one or both variables (e.g., predictor or criterion) to be correlated (Crocker & Algina, 1986). Restriction of range may also be caused in the observed scores if the predictor or criterion measures are too easy such that students earn relatively high scores (e.g., ceiling effect) or too difficult such that students earn very low scores (e.g., floor effect) (Algina & Crocker, 1986). Because of the potential for range restriction among middle grade struggling readers, it is important to examine the technical adequacy (e.g., reliability and validity) of ORF for struggling readers and students with disabilities at the beginning of the year as well as across progress monitoring time points.

Purpose of the Current Study and Research Questions

The purpose of the present study was to examine the reliability and validity of ORF among middle grade readers using equated passages. The study addressed the following research questions: (a) What is the **reliability** of the median and mean ORF scores among middle grade readers and are there significant differences in the **reliability** of the two scores when using equated scores across time points? (b) What is the magnitude of the relation (**validity**) between the median and mean ORF scores and external measures of reading fluency; are there significant differences in the relationship magnitude at the beginning of the year and across testing time points when using equated scores across time points? (c) What is the **reliability** of the mean and median ORF scores for struggling readers and does reliability vary across testing time points? (d) What is the magnitude of the relation (**validity**) between ORF and measures of reading proficiency when using the median versus mean scores for struggling readers; does it vary across testing time points? (e) Does the reliability and validity of the ORF median and mean scores differ between struggling reader and adequately developing reader groups?

Methods

Participants

School sites—This study was conducted in two large-urban cities in Texas. Approximately half of the sample was recruited from each site. The study participants were students from seven middle schools (grades 6–8). Students qualifying for reduced or free lunch ranged from 56% to 86% in the first site, and from 40% to 85% in the second site. Three of the seven schools were from a large urban district in one city with campus populations ranging from 500–1400 students. Four schools were from two medium size districts (school populations ranged in size from 633–1300) that drew both urban students from a nearby city and rural students from the surrounding areas. Of the seven schools, two were rated as recognized, four were rated as acceptable, and one school was rated academically unacceptable according to the state accountability system.

Students—The current study reports on 1,317 middle grade students from the seven schools during the 2006–2007 academic year. The sample includes 727 struggling readers and 590 adequate readers. Of the 1317 middle grade students, 52% were female, 38% were in sixth grade, 23% were in seventh grade, and 39% were in eighth grade. The sample is also ethnically diverse with African Americans comprising 40% of the sample, American Indians comprising less than 1%, Asians comprising less than 3%, Caucasians comprising 20%, and Hispanics comprising 37% of the total sample, which represents an oversampling of African Americans compared to Caucasians nationally or even in Texas.

Struggling readers were defined as students who either (a) failed the state reading achievement test (i.e., scores below 2100 points) (Texas Assessment of Knowledge and Skills; TAKS; Texas Educational Agency, 2008), or (b) performed within one-half of one standard error of measurement above the pass-fail cut-point on their first attempt in the spring of 2005 (i.e., scale scores ranging from 2100–2150 points) (see Vaughn et al., 2008 for more detail). In addition, students in special education who did not take TAKS-Reading but did take the State Developed Alternative Assessment (SDAA) reading test because of an exemption due to special education status were also defined as struggling readers.

Adequate readers were defined as students who obtained scale scores above 2150 on TAKS-Reading (performance above one-half of one standard error of measurement above the pass-fail cut-point) on their first attempt in the spring of 2005. A one-half of one standard error of measurement above the pass-fail cut-point was utilized to ensure that students who are highly likely to fail at future testing points due to measurement error associated with the test were considered struggling readers. Students were excluded from the study if: (a) they were enrolled in a special education life skills class with limited instructional time in general education; (b) their SDAA-Reading performance levels were equivalent to a 1st grade reading level or lower (i.e., nonreaders); (c) they presented a significant sensory disability, or (d) were classified as English as Second Language by their middle school. Because more than 80% of students pass TAKS-Reading, we randomly selected adequate readers within school (and grade) in relative proportion to the number of struggling readers. This resulted in a sample comprising 41% adequate readers at pretest.

Procedures

All participants were assessed by examiners who completed an extensive training program conducted by the investigators that focused on test administration, test scoring, and verification procedures for each measure included in the test battery. During the training examiners administered, scored, and verified many different samples to ensure accuracy. On the final day of training and prior to testing study participants, each examiner's testing performance was evaluated by the research team, only those examiners who demonstrated at least 95% accuracy for each test were permitted to evaluate study participants. All assessments were completed at the students' middle school in quiet locations designated by the school (i.e., library, unused classrooms, theatre, etc.). Following data collection, all student test packets were checked for accuracy of scoring. Packets were first scored by the examiner who tested the child; packets were then rescored by two examiners who did not test the child. Inter-scorer agreement exceeded 90 percent accuracy for all measures in the battery.

For reliability purposes, all students were assessed at five time points during the school year: (time 1) September, (time 2) November, (time 3) January, (time 4) March, and (time 5) May. At each testing time point (one through five), the test battery included three ORF passages. At each time point, the number of words read correctly was calculated. Raw scores were converted to equated scores in order to remove the effects of form. The median score and the mean score for the three passages was calculated using raw scores and equated scores.

For validity purposes we also administered Form A of the Test of Silent Reading Efficiency (TOSRE; Wagner, in press) and one grade level AIMSweb Maze CBM Reading Comprehension measure (Shinn & Shinn, 2002). At time points one and five, Form A of the Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1999) was also administered.

Measures

Measures to assess reliability

Passage Fluency (Francis, Barth, Reed, & Fletcher, 2008): The ORF consists of graded passages administered as short 1 minute probes to assess oral reading fluency. Across all time points, students read three passages for one minute each. For each of the three passages, the number of correct words read per minute (CWPM) was calculated. In order to more precisely estimate students' ORF abilities, raw scores (i.e., CWPM) were then converted to linearly equated scores on a story by story basis, within grade and time-point. The distribution of scores was statistically adjusted so that any given form has the same mean and standard deviation (Crocker & Algina, 1986). Adjustments were made relative to an anchor test, in this case the TOWRE Sight Word Efficiency. The equated scores eliminated differences between passages in mean differences and in within-passage variability at each assessment time point, but allowed differences over time and across grade in both mean performance and variability in performance. Therefore, differences in mean performance across time points and grades are preserved, and as a result, we can be sure that differences are not due to older grade students reading easier passages, or students reading difficult passages followed by easier passages later in the year, for example (see Vaughn et al., 2008 for more detail).

Linear equating was utilized because readability formulae imprecisely estimate passage equivalence (Frances et al., 2008). Consequently, changes in ORF scores across time may reflect differences in the scaling of forms themselves and may not reflect changes in students' achievement levels (Betts et al., 2009). For example, two passages (i.e., passage 1 and passage 2) may be of equivalent difficulty (mean 100 CWPM) but passage 1 may have a standard deviation of 10 words and passage 2 may have a standard deviation of 20 words. This variability ultimately impacts how ORF scores obtained from passage 1 and passage 2 can be interpreted. For instance, if a student reads passage 1 and passage 2 at a rate of 140 words correct per minute, one might be inclined to infer that the both passages were read at equivalent fluency rates. However, the student actually read passage 1 at a rate of 4 standard deviations above the mean and read passage 2 at a rate of two standard deviations above the mean. Thus the two scores are not comparable and do not represent the same unit of measurement (Betts et al., 2009).

Measures to assess convergent validity

Test of Sentence Reading Efficiency (TOSRE; Wagner, in press): The TOSRE, Form A, is a 3-minute, group-based assessment of reading fluency and comprehension. Students are presented with a series of short sentences, and are required to confirm the accuracy of each sentence. The mean intercorrelation across the five performances in the first year of the study was 0.79 for standard scores for students in Grade 6 and 0.92 for students in Grade 7–8. The standard score was the dependent measure utilized.

AIMSweb Maze CBM Reading Comprehension (Shinn & Shinn, 2002): The Maze CBM Reading Comprehension subtest is a 3-minute, group-based curriculum based assessment of fluency and comprehension. Students are presented with a 150–400 word passage and are required to identify the correct target among three choices for each omitted word in the passage. Howe and Shinn (2002) report a median test re-test reliability of 0.85 for students in Grade 6, 0.79 for students in grade 7, and 0.92 for students in grade. The raw score is the number of targets correctly identified in three-minutes and was the dependent measure utilized.

Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1999): At time point one the Sight Word Efficiency was administered. For the Sight Word Efficiency

subtest, the student was given a list of 104 real words and asked to read them as accurately and as quickly as possible. The raw score is the number of words read correctly within 45 seconds. Alternate forms and test retest reliability coefficients are at or above .90 for students in grades 6–8 (Torgesen, Wagner, & Rashotte, 1999). Standard scores were the dependent measure analyzed.

Analytic Approach—To compare the reliability of the median statistic with the reliability of the mean statistic we calculated test re-test reliability across alternate forms. Test re-test reliability across alternate forms requires constructing similar forms of a test and administering both forms to the same group of students across time points (Crocker & Algina, 1986). The higher the correlation between to two scores (i.e., correlation of equivalence), the higher the confidence one has in using the two scores interchangeably. For this study, alternate form correlations were calculated between time point 1 and 2, time point 1 and 3, time point 1 and 4, and time point 1 and 5 for both the mean and median. The duration of time between time points ranged from 5 to 11 weeks.

To determine whether the alternate form reliability of the median statistics was significantly higher than the alternate form reliability of the mean statistic, z-tests of the difference were calculated using a Bonferroni corrected alpha level to determine significance.

To examine the convergent validity of a given score, we computed correlations between ORF Mean and Median scores at time point 1 and external measures of reading fluency at time points 2–5. At time points 2–5 correlations were calculated among Mean and Median ORF scores and the TOSRE and AIMSweb Maze. Correlations among Mean and Median ORF scores with measures of reading fluency (e.g., TOWRE, TOSRE, and AIMSweb Maze) are also reported at time point 1.

To determine whether the Mean score was significantly more valid than the Median score, Hotelling Williams test for the difference between two dependent correlations (Tabachnick & Fidell, 2001) was calculated. Hotelling Williams test for the difference between two dependent correlations tests whether the correlation between the Mean score and external measures of fluency differs from the correlation between the Median score and external measures of fluency, where all variables are measured on the same students. This methodology was previously employed by Ardoin, Will, Suldo, Connell, Koenig, and Resetar et al. (2004) to determine if there was a significant difference in magnitude of relations among ORF median scores and measures of comprehension and single passage ORF scores and measures of comprehension.

Prior to analyses, we evaluated distributional data across all time points both statistically and graphically for skewness, kurtosis, and normality, with few violations. Across time points and variables, skewness ranged from −0.30 to 0.74 and kurtosis ranged from −0.07 to 1.05 for the full sample, skewness ranged from −0.004 to 0.74 and kurtosis ranged from 0.32 to 0.63 among adequate readers, and skewness ranged from −0.72 to 0.99 and kurtosis ranged from −0.08 to 1.5 among struggling readers. A total of 9 students were dropped because they did not complete the AIMS-web Maze task. For all remaining students in the study, no data was missing within testing time points and across testing time points.

Results

Reliability and validity of the median score

Table 1 shows the correlations among ORF mean, ORF median, and external measures of reading fluency (AIMSweb Maze CBM Reading Comprehension, TOSRE, and TOWRE) at time point 1. The ORF median score correlates moderately well with external measures of

reading fluency (range $r = 0.44–0.73$). Across time points and alternate forms, test re-test correlations are high (see Table 2). The test re-test correlation for the median score at time point 1 with time point 2 is 0.91; with time point 3 is 0.90; with time point 4 is 0.89; and with time point 5 is 0.87. The median ORF score possesses moderate convergent validity (see Table 3). The correlations among ORF median scores at time point 1 with AIMSweb Maze CBM Reading Comprehension range from $r = 0.57$ to 0.64 across the five time points. The correlations among ORF median score at time point 1 with TOSRE range from $r = 0.63$ to 0.68 across time points 1–5.

Reliability and validity of the mean score

Table 1 shows that at time point 1, the ORF mean score correlates moderately well with external measures of reading fluency (range r = 0.57 – 0.73). Across time points and alternate forms, the test re-test correlations are high (see Table 2). The test re-test correlations for the ORF mean score at time point 1 with the ORF mean score at time point 2 is 0.92, with time point 3 is 0.91, with time point 4 is 0.90, and with time point 5 is 0.88. The mean ORF score posses moderate convergent validity (see Table 3). Correlations among the ORF mean score at time point 1 and AIMSweb Maze CBM Reading Comprehension range from $r = 0.57$ to 0.63 across the five time points. The correlations among ORF mean score at time point 1 with the TOSRE range from $r = 0.64$ to 0.68.

Is the reliability and validity of ORF median and mean scores significantly different?

Testing the difference in test re-test reliability across alternate forms for the median and mean score indicated that the two statistics were not significantly different across all waves $p's > 0.025$ (see Table 2). Regarding convergent validity, Table 3 indicates that the magnitude of the relation among the Mean ORF score and external measures of reading fluency were generally equivalent to the magnitude of relation among the Median ORF and external measures of fluency across time points 2–5. The only exception to this is at time point 2, where the magnitude of the relation between the Mean score and TOSRE was greater than the Median score and TOSRE.

Reliability and validity of the median ORF score for struggling readers

Table 4 shows the correlations among ORF mean, ORF median, and external measures of reading fluency (AIMSweb Maze CBM Reading Comprehension, TOSRE, and TOWRE) at time point 1 for struggling readers. At time point 1, the ORF median score correlates moderately well with external measures of reading fluency (range $r = 0.44–0.77$). Across time points and alternate forms, the test re-test correlations are high (see Table 5). The test re-test correlation for the median score at time point 1 with time point 2 is 0.91; with time point 3 is 0.90; with time point 4 is 0.88; and with time point 5 is 0.85. The median score possesses moderate convergent validity (see Table 6). Among struggling readers, the correlation between the ORF median score at time point 1 with AIMSweb Maze CBM Reading Comprehension range from $r = 0.44$ to 0.51 across the five time points. The correlation between ORF median score at time point 1 with TOSRE range from $r = 0.43$ to 0.53 across time points 1–5.

Reliability and validity of the mean ORF score for struggling readers

Table 4 shows that among struggling readers, the ORF mean score correlates moderately well with external measures of reading fluency (range $r = 0.44–0.77$) at time point 1. Table 5 shows that the test retest reliability across alternate forms for the ORF mean score at time point 1 with the ORF mean score at time points 2–5 range from $r = 0.86$ to 0.92. Regarding convergent validity, the correlations among the mean ORF score at time point 1 and

AIMSweb Maze CBM Reading Comprehension range from 0.43 to 0.51 at time points 1–5 and with the TOSRE range from 0.43 to 0.54 among struggling readers (see Table 6).

Is the reliability and validity of the ORF median and mean score significantly different among middle grade struggling readers?

Testing the differences in test re-test reliability across alternate forms, for the median and mean ORF score indicated that the statistics were not significantly different among struggling readers, *z range* = −1.17 to 0 (see Table 5). Regarding convergent validity, the correlation among the median ORF score with external measures of reading fluency was not significantly different from the mean ORF score with external measures of reading fluency *p*'s > .05 (see Table 6).

Reliability and validity of the median ORF score for adequate readers

Table 7 shows the correlations among ORF mean, ORF median, and external measures of reading fluency (AIMSweb Maze CBM Reading Comprehension, TOSRE, and TOWRE) at time point 1 for adequate readers. At time point 1, the ORF median score correlates moderately well with external measures of reading fluency (range $r = 0.54–0.58$). Across time points, test alternate form correlations are high (see Table 8). Test re-test reliability across alternate forms for the median score at time point 1 with time point 2 is 0.87; with time point 3 is 0.86; with time point 4 is 0.85; and with time point 5 is 0.83. The median score possesses moderate convergent validity (Table 9). Among adequate readers, the correlation between the ORF median score at time point 1 with AIMSweb Maze CBM Reading Comprehension range from $r = 0.55$ to 0.64 across the five time points. The correlation between ORF median score at time point 1 with TOSRE range from $r = 0.58$ to 0.63 across time points 1–5.

Reliability and validity of the mean ORF score for adequate readers

Table 7 shows that among adequate readers, the ORF mean score correlates moderately well with external measures of reading fluency (range $r = 0.55–0.59$) at time point 1. Table 8 shows that test re-test reliability across alternate forms for the ORF mean score at time point 1 with the ORF mean score at time points 2–5 range from $r = 0.83$ to 0.88. Regarding convergent validity, the correlations among the mean ORF score at time point 1 and AIMSweb Maze CBM Reading Comprehension range from 0.55 to 0.65 at time points 1–5 and with the TOSRE range from 0.59 to 0.64 among adequate readers (see Table 9).

Is the reliability and validity of the ORF median and mean score significantly different among middle grade adequate readers?

Testing the differences in test re-test reliability for the median and mean ORF score indicated that in the scores were not significantly different among adequate readers, $z = -0.73$ to 0 (see Table 8). Regarding convergent validity, the general pattern of results suggests that the magnitude of the relations among the Median ORF score and external measures of reading fluency are not significantly different that the magnitude of relations among the Mean ORF score and external measures of fluency. Exceptions to this pattern are at time points 2, where the magnitude of relations among the Mean ORF score and external measures of fluency is greater in magnitude than that of the Median ORF score and external measures of reading fluency (see Table 9).

Is the reliability and validity of the ORF median and mean score for struggling readers significantly different than that of adequate readers?

Test of the difference in test re-test reliability for the median score among struggling and adequate readers indicated that the reliability of the median ORF score among struggling

readers is greater than that of adequate readers across time points 2–4 (see Table 10). Also, the test re-test reliability of the mean ORF score among struggling readers are significantly higher than that of the mean ORF score among adequate readers at time point 3. Regarding convergent validity, correlations of the median and mean score with AIMSweb Maze CBM Reading Comprehension and TOSRE were significantly greater in magnitude among adequate readers than struggling readers across time points (see Table 11).

Discussion

The present study examined the reliability and validity of ORF among middle grade readers using linearly equated passages. The results reveal that among middle grade readers, ORF measures are highly reliable across time points and possess moderate convergent validity. Among middle grade readers with reading disabilities and difficulties, ORF measures are also highly reliable and moderately valid. Among students without reading disabilities and reading difficulties (e.g., adequately developing middle grade readers), a similar pattern of findings was obtained.

This study also compared the reliability and validity of the ORF median and ORF mean score for struggling readers with that of adequate readers to determine if the reliability and validity was significantly different between the two groups. Results indicate that the alternate form reliability of ORF mean and median scores were significantly greater among struggling readers. Correlations among ORF mean and median with external measures of reading fluency were moderate for both groups of readers, but were significantly greater in magnitude among adequately developing readers.

Results of this study provide further evidence supporting the use of the median score among middle grade readers. The use of the median score evolved as a method of dealing with the tendency of passages, presumed equivalent in difficulty, to yield varying estimates of reading fluency ability. The primary advantage of using the median score is that is based on a single story and requires no further manipulation for interpretation. The major disadvantage of using the median score is that it disregards two key data points (i.e., the low score and high score). Variable scores (low, middle, and high) frequently resulted because the administered passages were not of equivalent difficulty. However, variability in student performance is not simply a matter of text inequality. Measurement error also prevents an observed score from a single test, or even scores from multiple tests administered at a simple time point, from perfectly capturing ability (Fletcher, Denton, & Francis, 2005).

The mean score does not eliminate the measurement issues involved in quantifying reading ability. However, it would seem that the mean score would likely increase the reliability of the fluency estimate since the random effects that represent measurement error are averaged together. A partial cancellation (averaging) of effects of measurement can be expected, which would seem to lead to a more reliable estimate of reading ability. However, results of this study indicate that when using linearly equated scores, both the median and mean score possess high test-retest reliability and moderate convergent validity.

Finally, in middle school, oral reading fluency measures are primarily used with students in special education. Our results indicate that the test retest reliability of the median score was not significantly different from the mean score. Additionally, the correlation of the median with external measures of fluency was not significantly different than the mean score. Interestingly, when comparing the alternate form reliability between struggling readers and adequate readers, the reliability of the mean and median ORF scores are significantly higher among struggling readers. However, when comparing the validity of the ORF median and

mean scores, the convergent validity is stronger in magnitude among adequately developing readers.

Limitations

This study converted raw scores to equated scores in order to ensure that the measurement of oral reading fluency ability was not unduly influenced by differences in text type, text difficulty, or order of administration. The use of equated scores represents a departure from traditional administration procedures but ensures that the underlying assumption of passage equivalence is met (Francis et al., 2008). Further, the extent to which the findings of this study can be generalized may depend upon whether parallel forms are available for administration and whether raw scores can be converted to equated scores.

Further discussion of the sample is also noteworthy. Specifically, Caucasians and Hispanics are under-represented and African Americans are over-represented when compared to the proportion of Caucasians, Hispanics, and African Americans ages 10–16 years in the state of Texas. For instance, in the state of Texas, 38% of children ages 10–16 years were Caucasian (i.e., Anglos), 13–14% African American, and 45% Hispanic (Texas State Data Center and Office of the State Demographer, 2009). In our sample, 20% of students were Caucasian, 37% Hispanic, and 40% African American. Such sample selection procedures may have contributed to restriction of range. However, the amount of variability observed among African Americans, Caucasians, and Hispanics were similar for the mean and median score. For the full sample of students, the standard deviations ranged from 35.4 to 38.4 for the mean score and 35.4 to 38.7 for the median score. Among African Americans the standard deviations ranged from 35.4 to 37.9 for the mean score and 35.7 to 38.3 for the median score. Among Caucasians the standard deviations ranged from 36.5 to 39.3 for the mean score and 36.2 to 40.1 for the median score. Among Hispanics the standard deviations ranged from 31.8 to 35.0 for the mean score and 31.6 to 35.1 for the median score. In addition, across variables and time points, skewness and kurtosis were not significant within the full sample, adequate readers, or struggling readers. Altogether, this suggests that sample selection procedures did not result in restriction of range for passage fluency (e.g., mean or median scores) or the external measures of fluency (e.g., TOSRE and AIMS-Web Maze Reading Comprehension).

Implications for Practice

This study expands the small body of psychometric literature on ORF reading probes for middle grade readers and the large body of literature on ORF generally, showing that these ORF passages are highly reliable and moderately valid for middle grade readers. Of the two scores evaluated (i.e., mean and median passage scores), the reliability and validity of the mean and median are equivalent. Thus, when administering CBM reading probes to middle grade readers, the examiner may select either the mean or median score.

Future Directions

The moderate validity of oral reading fluency CBM suggests that these measures might represent a reliable and valid indicator of academic performance among middle grade readers. However, to substantiate such a claim, it is necessary to first examine the relation between ORF and measures of reading comprehension and to then determine how much variance ORF accounts for in reading comprehension ability. Addressing these questions could determine whether CBM oral reading probes represents reliable and valid indicators of overall reading proficiency among middle grade readers.

The moderate validity of the ORF does not mean that teachers require little or no other data sources to guide their instructional decision making. In addition to ORF, we would expect

that teachers would be interested in how students' demonstrate understanding and learning from text, how motivated they are to read, and the extent to which they demonstrate text understanding through oral discourse and in their writing.

Acknowledgments

This paper was supported in part by a grant from the National Institute of Child Health and Human Development, P50 HD052117, Texas Center for Learning Disabilities.

References

Betts J, Pickart M, Heistad D. An investigation of the psychometric evidence of CBM-R passage equivalence: Utility of readability statistics and equating for alternate forms. Journal of School Psychology. 2009; 47:1–17.

Crocker, LM.; Algina, J. Introduction to classical and modern test theory. Belmont, CA: Wadsworth Group/Thomas Learning; 1986.

Deno S. The nature and development of curriculum-based measurement. Preventing School Failure. 1992; 36:5–10.

Deno S, Mirkin P, Chaing B. Identifying valid measures of reading. Exceptional Children. 1982; 49:524–597.

Espin C, Foegen A. Validity of general outcome measures for predicting secondary students' performance on content-area tasks. Exceptional Children. 1996; 62:497–514.

Espin C, Wallace T, Lembke E, Campbell H, Long JD. Creating a progress-monitoring system in reading for middle-school students: Tracking progress toward meeting high-stakes standards. Learning Disability Research and Practice. 2010; 25:60–75.

Fletcher JM, Denton C, Francis D. Validity of alternative approaches for the identification of learning disabilities: Operationalizing unexpected underachievement. Journal of Learning Disabilities. 2005; 38:545–552. [PubMed: 16392697]

Francis, DJ.; Barth, AE.; Reed, D.; Fletcher, JM. Texas Middle School Fluency Assessment. University of Houston; 2008.

Francis DJ, Santi KL, Barr C, Fletcher JM, Varisco A, Foorman BR. Form effects on the estimation of students' oral reading fluency using DIBELS. Journal of School Psychology. 2008; 46:315–342. [PubMed: 19083362]

Fuchs L, Deno D, Mirkin P. Effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. American Educational Research Journal. 1984; 21:449–460.

Fuchs L, Fuchs D. Effects of systematic formative evaluation on student achievement. Exceptional Children. 1986; 53:199–207. [PubMed: 3792417]

Fuchs L, Fuchs D, Compton D. Monitoring early reading development in first grade: Word identification fluency versus nonsense word fluency. Exceptional Children. 2004; 71:7–21.

Fuchs L, Fuchs D, Hamlett C, Ferguson C. Effects of expert system consultation within curriculum-based measurement using a reading maze task. Exceptional Children. 1992; 58:436–450.

Fuchs LS, Fuchs D, Maxwell L. The validity of informal reading comprehension measures. Remedial and Special Education. 1988; 9:20–28.

Hintze JM, Silberglitt B. A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. School Psychology Review. 2005; 34:372–386.

Jenkins JR, Fuchs LS, van den Broek P, Espin CA, Deno S. Sources of individual differences in reading comprehension and reading fluency. Journal of Educational Psychology. 2003; 95:719–729.

Jenkins JR, Jewell M. Examining the validity of two measures for formative teaching: Reading aloud and maze. Exceptional Children. 1993; 59:421–432.

Marston, D. A curriculum-based measurement approach to assessing academic performance: What it is and why do it. In: Shinn, M., editor. Curriculum-based measurement: Assessing special children. New York: Guilford Press; 1989. p. 18-78.

Mirkin PK, Deno S, Tindal G, Kuehnle K. Frequency of measurement and data utilization strategies as factors in standardized behavioral assessment of academic skill. Journal of Behavioral Assessment. 1982; 4:361–370.

McGlichey MT, Hixson MD. Using curriculum based measurement to predict performance on state assessments in reading. School Psychology Review. 2004; 33:193–203.

Kranxler J, Brownell MT, Miller MD. The construct validity of curriculum-based measurement of reading: An empirical test of a plausible rival hypothesis. Journal of School Psychology. 1998; 36:399–415.

Powell-Smith, K.; Stewart, LH. The use of curriculum based measurement on the reintegration of students with mild disabilities. In: Shinn, MR., editor. Advanced applications of curriculum-based measurement. New York: Guildford Press; 1998. p. 254-307.

Reschly A, Busch T, Betts J, Deno S, Long J. Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. Journal of School Psychology. 2009; 47:427–469. [PubMed: 19808123]

Shinn, MR. Curriculum-based measurement: Assessing special children. New York: Guilford Press; 1989.

Shinn MR, Good RH, Knutson N, Tilly WD, Collins VL. Curriculum-based measurement or oral reading fluency: A confirmatory analysis of its relation to reading. School Psychology Review. 1992; 21:459–479.

Shinn MR, Powell-Smith KA, Good RH, Baker S. The effects of reintegration into general education reading instruction for students with mild disabilities. Exceptional Children. 1997; 64:59–80.

Shinn, MR.; Shinn, MM. AIMSweb training workbook. Eden Prairie, MN: Edformation, Inc; 2002.

Silberglitt B, Burns MK, Madyn NH, Lail KE. Relationship of reading fluency assessment data with state accountability test scores: A longitudinal comparison of grade levels. Psychology in the Schools. 2006; 43:527–535.

Stage SA, Jacobson MD. Predicting student success on a state-mandated performance-based assessment using oral reading fluency. School Psychology Review. 2001; 30:407–419.

Stecker PM, Fuchs LS. Effecting superior achievement using curriculum-based measurement: the importance of individual progress monitoring. Learning Disabilities Research and Practice. 2000; 15:128–134.

Stecker PM, Fuchs LS, Fuchs D. Using curriculum-based measurement to improve student achievement: Review of Research. Psychology in Schools. 2005; 42:795–819.

Texas Educational Agency. TAKS: Texas Assessment of Knowledge and Skills. 2004a. Downloaded from http://www.tea.state.tx.us/student.assessment/taks/booklets/reading/g5e.pdf on December 19, 2007

Institute for Demographic and Socioeconomic Research. Estimates of the Population by Age, Sex, and Race/Ethnicity for July 1, 2009 for the State of Texas. 2009. Downloaded from http://txsdc.utsa.edu/tpepp/txpopest.php on January 21, 2011

Ticha R, Espin C, Wayman MW. Reading progress monitoring for secondary-school students: Reliability, validity, and sensitivity to growth of reading aloud and maze selection measures. Learning Disabilities Research and Practice. 2009; 24:132–142.

Tindal, G. Evaluating the effectiveness of educational programs at the systems level using curriculum-based measurement. In: Shinn, M., editor. Curriculum-based assessment: Assessing special children. New York: Guilford Press; 1989. p. 202-238.

Torgesen, J.; Wagner, R.; Rashotte, C. Test of Word Reading Efficiency. Austin, TX: Pro-Ed; 1999.

Wagner, R. Test of Sentence Reading Efficiency. Austin, TX: Pro-Ed; (in press)

Wayman C, Wallace T, Wiley H, Ticha R, Espin C. Literature synthesis on curriculum-based measurement in reading. Journal of Special Education. 2007; 41:85–120.

Yovanoff P, Duesbery L, Alonzo J, Tindal G. Grade-level invariance of a theoretical causal structure predicting reading comprehension with vocabulary and oral reading fluency. Educational Measurement: Issues and Practice. 2005; 24:4–12.

Table 1

Correlation Matrix of Time Point 1 Variables

Variable	1.	2.	3.	4.	5.
1. Mean Passage Fluency	1.0				
2. Median Passage Fluency	.99	1.0			
3. AIMSweb Maze CBM Reading Comprehension	.57	.57	1.0		
4. TOSRE	.67	.66	.53	1.0	
5. TOWRE	.73	.73	.44	.59	1.0
Mean	127.7	127.9	173.7	90.4	96.8
Standard Deviation	35.4	35.5	64.0	15.0	12.1

Note. n = 1317. Mean Passage Fluency = Mean score of three one minute fluency probes; Median Passage Fluency = median score of three one minute fluency probes; TOSRE = Test of Sentence Reading Efficiency; TOWRE = Test of Word Reading Efficiency, Sight Word Efficiency subtest.

Table 2

Test Re-test Reliability across Alternate Forms for ORF Median and Mean Score across Time Points and z-score test of the Difference in Test Re-test Reliability across Alternate Forms for Testing Time Points 1–5

	Test Re-test Reliability across Alternate Forms for the ORF Median and Mean Score (Time Point 1 with Time Points 2–5)			
	Time Point 1 – Time Point 2	**Time Point 1 – Time Point 3**	**Time Point 1 – Time Point 4**	**Time Point 1 – Time Point 5**
Median	.91	.90	.89	.87
Mean	.92	.91	.90	.88
	Testing the Magnitude of the Difference between the Test Re-test Reliability ORF Median and Mean Score (Time Point 1 with Time Points 2–5)			
	Time Point 1 – Time Point 2	Time Point 1 – Time Point 3	Time Point 1 – Time Point 4	Time Point 1 – Time Point 5
Median vs. Mean	−1.58 [a]	−1.42 [a]	1.29 [a]	−1.20 [a]

Note. $n = 1317$.

[a] alternate form reliability of ORF Median Score is statistically higher than ORF Mean Score;

[b] alternate form reliability of ORF Mean Score is statistically higher than Median Score. Alpha per comparison = 0.05/2 = 0.025. Critical $z = 2.33$.

Table 3

Correlations of ORF Median and Mean Scores at Time Point 1 with External Measures of Reading Fluency across Time Points 2–5 and Test of the Difference the Magnitude of Median and Mean Correlations at Time Point 1 with External Measures of Reading Fluency Across Points 2–5

Correlations of ORF Median and Mean Scores at Time Point 1 with External Measures of Reading Fluency at Time Points 2–5								
	Time Point 2		**Time Point 3**		**Time Point 4**		**Time Point 5**	
	Maze	**TOSRE**	**Maze**	**TOSRE**	**Maze**	**TOSRE**	**Maze**	**TOSRE**
Median	.57	.68	.63	.63	.63	.65	.57	.65
Mean	.57	.68	.63	.64	.63	.66	.57	.66
	Testing the Magnitude of the Difference between the ORF Mean and Median Correlations at Time Point 1 with External Measures of Reading Fluency across Points 2–5							
	Time Point 2		Time Point 3		Time Point 4		Time Point 5	
	Maze	TOSRE	Maze	TOSRE	Maze	TOSRE	Maze	TOSRE
Median vs. Mean	1.85	2.16 *b*	.44	1.86	1.67	1.51	.05	1.03

Note. $n = 1317$. $p < 0.05$. Maze = AIMS-Web Maze CBM Reading Comprehension; TOSRE = Test of Sentence Reading Efficiency.

a validity of Median ORF Score is greater in magnitude than ORF Mean Score;

b validity of the ORF Mean Score is greater in magnitude than ORF Median Score.

Table 4

Correlation Matrix of Time Point 1 Variables among Struggling Readers

Variable	1.	2.	3.	4.	5.
1. Mean Passage Fluency	1.0				
2. Median Passage Fluency	.99	1.0			
3. AIMSweb Maze CBM Reading Comprehension	.44	.44	1.0		
4. TOSRE	.54	.54	.33	1.0	
5. TOWRE	.77	.77	.36	.55	1.0
Mean	113.3	112.9	153.2	83.3	92.3
Standard Deviation	32.6	32.5	58.1	12.6	10.8

Note. $n = 727$. Mean Passage Fluency = Mean score of three short one minute fluency probes; Median Passage Fluency = median score of three short one minute fluency probes; TOSRE = Test of Sentence Reading Efficiency; TOWRE = Test of Word Reading Efficiency, Sight Word Efficiency subtest.

Table 5

Test Re-test Reliability across Alternate Forms for ORF Median and Mean Score and z-score test of the Difference in A Reliability of across Testing Time Points 1–5 for Struggling Readers.

	Test Re-test Reliability across Alternate Forms for the ORF Median and Mean Score (Time Point 1 with Time Points 2–5)			
	Time Point 1 – Time Point 2	**Time Point 1 – Time Point 3**	**Time Point 1 – Time Point 4**	**Time Point 1 – Time Point 5**
Median	.91	.90	.88	.85
Mean	.92	.90	.88	.86
	Testing the Magnitude of the Difference between the Test Re-test Reliability ORF Median and Mean Score (Time Point 1 with Time Points 2–5)			
	Time Point 1 – Time Point 2	Time Point 1 – Time Point 3	Time Point 1 – Time Point 4	Time Point 1 – Time Point 5
Median vs. Mean	−1.17	0	0	−.71

Note. $n = 727$.

[a] alternate form reliability of ORF Median Score is statistically higher than ORF Mean Score;

[b] alternate form reliability of ORF Mean Score is statistically higher than Median Score. Alpha per comparison = 0.05/2 = 0.025. Critical $z = 2.33$.

Table 6

Correlations of ORF Median and Mean Scores at Time Point 1 with External Measures of Reading Fluency across Time Points 2–5 and Test of the Difference the Magnitude of Median and Mean Correlations at Time Point 1 with External Measures of Reading Fluency Across Points 2–5among Struggling Readers

	Correlations of ORF Median and Mean Scores at Time Point 1 with External Measures of Reading Fluency at Time Points 2–5							
	Time Point 2		**Time Point 3**		**Time Point 4**		**Time Point 5**	
	Maze	**TOSRE**	**Maze**	**TOSRE**	**Maze**	**TOSRE**	**Maze**	**TOSRE**
Median	.51	.53	.51	.43	.48	.48	.45	.51
Mean	.51	.53	.51	.43	.50	.48	.45	.51
	Testing the Magnitude of the Difference between the ORF Mean and Median Correlations at Time Point 1 with External Measures of Reading Fluency across Points 2–5							
	Time Point 2		Time Point 3		Time Point 4		Time Point 5	
	Maze	TOSRE	Maze	TOSRE	Maze	TOSRE	Maze	TOSRE
Median vs. Mean	.36	.98	.84	.67	.31	.89	.97	.85

Note. $n = 727$. $p < 0.05$. Maze = AIMS-Web Maze CBM Reading Comprehension; TOSRE = Test of Sentence Reading Efficiency.

[a] validity of Median ORF Score is greater in magnitude than ORF Mean Score;

[b] validity of the ORF Mean Score is greater in magnitude than ORF Median Score.

Table 7

Correlation Matrix of Time Point 1 Variables among Adequate Readers

Variable	1.	2.	3.	4.	5.
1. Mean Passage Fluency	1.0				
2. Median Passage Fluency	.99	1.0			
3. AIMSweb Maze CBM Reading Comprehension	.56	.55	1.0		
4. TOSRE	.59	.58	.55	1.0	
5. TOWRE	.55	.54	.33	.40	1.0
Mean	145.9	145.9	199.1	99.1	102.3
Standard Deviation	30.0	30.1	62.9	13.0	11.4

Note. n = 590. Mean Passage Fluency = Mean score of three short one minute fluency probes; Median Passage Fluency = median score of three short one minute fluency probes; TOSRE = Test of Sentence Reading Efficiency; TOWRE = Test of Word Reading Efficiency, Sight Word Efficiency subtest.

Table 8

Test Re-test Reliability across Alternate Forms for ORF Median and Mean Score across Time Points and z-score test of the Difference in Test Re-test Reliability across Alternate Forms for Testing Time Points 1–5 for Adequate Readers.

	Test Re-test Reliability across Alternate Forms for the ORF Median and Mean Score (Time Point 1 with Time Points 2–5)			
	Time Point 1 – Time Point 2	**Time Point 1 – Time Point 3**	**Time Point 1 – Time Point 4**	**Time Point 1 – Time Point 5**
Median	.87	.86	.84	0.83
Mean	.88	.87	.85	0.83
	Testing the Magnitude of the Difference between the Test Re-test Reliability ORF Median and Mean Score (Time Point 1 with Time Points 2–5)			
	Time Point 1 – Time Point 2	Time Point 1 – Time Point 3	Time Point 1 – Time Point 4	Time Point 1 – Time Point 5
Median vs. Mean	−.73	−.69	0	0

Note. $n = 590$.

[a] alternate form reliability of ORF Median Score is statistically higher than ORF Mean Score;

[b] alternate form reliability of ORF Mean Score is statistically higher than Median Score. Alpha per comparison = 0.05/2 = 0.025. Critical $z = 2.33$.

Table 9

Correlations of ORF Median and Mean Scores at Time Point 1 with External Measures of Reading Fluency across Time Points 2–5 and Test of the Difference the Magnitude of Median and Mean Correlations at Time Point 1 with External Measures of Reading Fluency Across Points 2–5 for Adequate Readers

	Correlations of ORF Median and Mean Scores at Time Point 1 with External Measures of Reading Fluency at Time Points 2–5							
	Time Point 2		**Time Point 3**		**Time Point 4**		**Time Point 5**	
	Maze	**TOSRE**	**Maze**	**TOSRE**	**Maze**	**TOSRE**	**Maze**	**TOSRE**
Median	.64	.63	.60	.61	.62	.62	.56	.60
Mean	.65	.64	.61	.62	.62	.63	.55	.60
	Testing the Magnitude of the Difference between the ORF Mean and Median Correlations at Time Point 1 with External Measures of Reading Fluency across Points 2–5							
	Time Point 2		Time Point 3		Time Point 4		Time Point 5	
	Maze	TOSRE	Maze	TOSRE	Maze	TOSRE	Maze	TOSRE
Median vs. Mean	2.06 *b*	2.33 *b*	.44	1.37	1.05	1.07	.28	.76

Note. $n = 590$. $p < 0.05$. Maze = AIMS-Web Maze CBM Reading Comprehension; TOSRE = Test of Sentence Reading Efficiency.

a validity of Median ORF Score is greater in magnitude than ORF Mean Score;

b validity of the ORF Mean Score is greater in magnitude than ORF Median Score.

Table 10

Comparing the Alternate Form Reliability Correlations for ORF Median and Mean Score across Time Points for Struggling and Adequate Readers and z-score test of the Difference in Alternate Form Reliability of across Testing Time Points 1–5 for Struggling and Adequate Readers.

Alternate Form Reliability Correlations of the ORF Median and Mean Score (Time Point 1 with Time Points 2–5) Comparing Struggling and Adequate Readers								
	Time Point 1 – Time Point 2		**Time Point 1 – Time Point 3**		**Time Point 1 – Time Point 4**		**Time Point 1 – Time Point 5**	
	Struggling Readers	**Adequate Readers**	**Struggling Readers**	**Adequate Readers**	**Struggling Readers**	**Adequate Readers**	**Struggling Readers**	**Adequate Readers**
Median	.91	.87	.90	.86	.88	.85	.85	.83
Mean	.92	.88	.90	.87	.88	.85	.86	.83
Testing the Magnitude of the Difference between the ORF Median and Mean Alternate Form Correlations across Time Points (Time Point 1 with Time Points 2–5)								
	Time Point 1 – Time Point 2		Time Point 1 – Time Point 3		Time Point 1 – Time Point 4		Time Point 1 – Time Point 5	
Median (Struggling vs. Adequate Readers)	3.5 [a]		3.2 [a]		2.2		1.2	
Mean (Struggling vs. Adequate Readers)	1.1		2.5 [c]		2.2		1.9	

Note. n = 727 Struggling Readers. n = 590 Adequate Readers.

[a] alternate form reliability of ORF Median Score among struggling readers is statistically higher than ORF Median Score among adequate readers;

[b] alternate form reliability of ORF Median Score among adequate readers is statistically higher than Median Score among struggling readers.

[c] alternate form reliability of ORF Mean Score among struggling readers is statistically higher than ORF Mean Score among adequate readers;

[d] alternate form reliability of ORF Mean Score among adequate readers is statistically higher than Mean Score among struggling readers. Alpha per comparison = 0.05/2 = 0.025. Critical z = 2.33.

Table 11

Correlations of the Median Passage Score and Mean Passage Score with External Measures of Reading Fluency Across Five Waves for Struggling and Adequate Readers and Test of the Difference in Correlations Across Waves (z Scores) for the Two Groups

	Correlations of ORF Median and Mean Scores at Time Point 1 with External Measures of Reading Fluency at Time Points 2–5															
	Time Point 2				**Time Point 3**				**Time Point 4**				**Time Point 5**			
	Maze		**TOSRE**		**Maze**		**TOSRE**		**Maze**		**TOSRE**		**Maze**		**TOSRE**	
	SR	**AR**	**SR**	**AR**	**SR**	**AR**	**SR**	**AR**	**SR**	**AR**	**SR**	**AR**	**SR**	**AR**	**SR**	**AR**
Median	.51	.64	.53	.63	.51	.60	.43	.61	.48	.62	.48	.62	.45	.56	.51	.60
Mean	.51	.65	.53	.64	.51	.59	.43	.62	.50	.62	.48	.63	.45	.55	.51	.60
	Testing the Magnitude of the Difference between the ORF Mean and Median Correlations at Time Point 1 with External Measures of Reading Fluency across Points 2–5															
	Time Point 2				Time Point 3				Time Point 4				Time Point 5			
	Maze		TOSRE		Maze		TOSRE		Maze		TOSRE		Maze		TOSRE	
Median (Struggling vs. Adequate Readers)	−3.51 [b]		−2.72 [b]		−2.35 [b]		−4.48 [b]		−3.64 [b]		−3.64 [b]		−2.67 [b]		−2.35 [b]	
Mean (Struggling vs. Adequate Readers)	−3.83 [b]		−3.03 [b]		−2.07		−4.77 [b]		−3.16 [b]		−3.64 [b]		−2.41 [b]		−2.35 [b]	

Note. $n = 590$. $p < 0.05$. Maze = AIMS-Web Maze CBM Reading Comprehension; TOSRE = Test of Sentence Reading Efficiency.

[a] validity of Median ORF Score among struggling readers is greater in magnitude than ORF Median Score among adequate readers;

[b] validity of the ORF Median Score among adequate readers is greater in magnitude than ORF Median Score among struggling readers.

[c] validity of the ORF Mean Score among struggling readers is greater in magnitude than the ORF Mean Score among adequate readers;

[d] validity of the ORF Mean Score among adequate readers is greater in magnitude than the ORF Mean Score among struggling readers. Alpha per comparison = 0.05/2 = 0.025. Critical $z = 2.33$.