Technical Report # 22

Analysis of Reading Fluency and Comprehension Measures for Third Grade Students

Leanne Ketterlin Geller

Gerald Tindal

University of Oregon

Published by

Behavioral Research and Teaching
University of Oregon • 175 Education
5262 University of Oregon • Eugene, OR 97403-5262
Phone: 541-346-3535 • Fax: 541-346-5689
http://brt.uoregon.edu

Abstract

This technical report outlines the results of a correlational study of an Oral Reading Fluency (ORF) measure, a Reading Comprehension measure, a Vocabulary measure, and a statewide, large-scale reading assessment. The effects of school income level, gender, ethnicity, Special Education status, and English Language Learner status are also considered. For the ORF, statistically significant differences were found in all of the demographic comparisons. For the Reading Comprehension and Vocabulary tests, statistically significant differences were found in all comparisons except for gender. A correlational analysis indicated a strong correlation between the ORF measure and the Vocabulary test, a moderately strong correlation between the Vocabulary test and the statewide reading assessment, and weak to moderate correlations between all other measures. A regression analysis indicated that ORF, Reading Comprehension and Vocabulary measures predict 25% of the variance in statewide assessment scores.

Introduction

Third grade marks a pivotal point in reading instruction. Beginning with this grade, instruction focuses on reading to learn instead of learning to read. To learn from reading, however, students must comprehend the text. Thus, the demands on the learner shift from word recognition and decoding to gaining meaning. Therefore, students must develop proficiency in these prerequisite skills in 3rd grade; as a consequence, monitoring progress toward proficiency in reading is essential for making appropriate and timely instructional and programmatic decisions. Early detection of reading problems allows teachers and parents to target instruction and provide interventions to support the learner's needs.

This report describes the findings from a district-wide assessment system designed to monitor the competencies of 3rd grade students in oral reading fluency, reading comprehension, and vocabulary. Statewide assessment data is used to examine the relationship between each skill and summative evaluation of reading proficiency.

Methods

In this section, we describe the setting and subjects, measurement development, research procedures, and data analyses.

Setting and Subjects

This study was conducted during spring 2003 in 29 elementary schools in an urban school district in the Pacific Northwest. Participants include 1,219 third-grade students. Sixty-six student records were removed from the analyses due to missing data or anomalies with the data. Three records were removed due to out of level testing. Of the remaining 1,153 students, 182 students received special education services and 25 were classified as English language learners. The demographic characteristics of 3rd graders in the district mirrors that of the state with 788

Caucasian students, 82 Hispanic students, 57 Asian students, 36 African American students, 27 Native American students, and 103 with other ethnicities; information was unavailable on the remaining 60 students. Approximately equal numbers of male and female students participated, *n* = 564 and *n* =588, respectively.

Design and Operational Procedures

Each participant read one passage orally and completed seven comprehension questions relating to the story. Additionally, each student completed a 25-item vocabulary task. State assessment tests were administered to every student during the statewide testing window in April and May, 2003.

Measurement/Instrument Development

The blocking factors in this report include individual educational classification, such as receiving special education services, English language learner, or receiving Title services; additionally, gender and ethnicity are included. The dependent variables are oral reading fluency, reading comprehension, and vocabulary as measured by district-administered assessments.

Oral Reading Fluency

Oral reading fluency (ORF) was measured using district passages of approximately 250 words using standardized administration and scoring procedures. Each student was individually administered the same narrative passage to read orally by a trained tester. Students were provided with the pronunciation of proper names prior to being asked to read. A score for the ORF task was calculated by timing the student as they read the passage aloud for one minute. The number of errors was subtracted from the total words read to determine the number of words read correctly per minute.

Reading Comprehension

Immediately following administration of the ORF, students read the remainder of the story. Reading comprehension was measured using two constructed response items and five selected response questions. A district-wide team of teachers, psychologists, and staff development specialists developed all comprehension questions. Administration and scoring were standardized.

The constructed-response items were administered orally and the tester recorded student responses to each question. The administrator was allowed to repeat any item one time to each student. To score the response, the administrator used a 2-point, criterion-referenced scale. The maximum score was four points. For selection-response items, students chose the best answer from four options to complete a statement with items scored dichotomously (correct and incorrect). The maximum score was 5 points.

Vocabulary

A 25-item vocabulary test was administered. Students were instructed to read each word printed in bold-face type and then select an answer from a set of three options reflecting the most similar in meaning to the bold-faced word. Student selection was made by filling in a circle next to the answer choice. One point was awarded for each correct answer with the maximum possible points of 25.

Statewide Assessment in Reading

As part of the state testing program, each student was administered a multiple-choice assessment in reading with seven domains: word meaning, locating information, literal comprehension, inferential comprehension, evaluative comprehension, literary forms, and literary elements.

Data Preparation and Analysis

For each measure, descriptive statistics were computed, including an option analysis for the selection-response questions. Item difficulty was calculated by determining the percentage of students correctly answering each item. Item discrimination was calculated by gender, ethnicity, and educational classification using *t*-tests and analysis of variance statistical techniques. Item Response Theory (IRT) was used to examine item functioning. Finally, to determine the relationship between each dependent measure, correlational analyses were conducted using a statistically significant alpha level of $p < .0083$ to adjust for the additional measures, and multiple regression was also used to understand how the measures fit together.

Results

Oral Reading Fluency

The mean score for oral reading fluency was 105.9 words read correctly per minute; the standard deviation was 43.4. Descriptive statistics for oral reading fluency by ethnicity and educational classification are presented in Table 1.

Table 1

Descriptive statistics for oral reading fluency by gender, ethnicity, and educational classification.

Group	*n*	Mean	Standard Deviation
	Gender		
Male	564	99.8	42.0
Female	588	111.7	44.0
	Ethnicity		
Caucasian	788	109.3	42.7
Hispanic	82	88.9	36.7
Asian	57	119.2	47.9
African American	36	90.4	42.7
Native American	27	82.7	36.9
Other	103	105.3	46.4
	Educational Classification		
General education	723	110.4	38.7
Special education	182	70.3	40.8
English-language learner	25	77.0	42.9

Statistically significant differences in mean scores were observed for gender, $t(1150) = -4.69$, $p < .05$: girls' performed significantly better than boys'.

Statistically significant differences also were observed for ethnicity, $F(5,1087) = 7.17$, $p < .05$. Levene's test for homogeneity of variance was statistically significant; therefore post hoc analyses were conducted using Tomhane's 12 adjustments to determine the differences between

groups. Native Americans performed statistically lower than Asian and Caucasian students; Hispanic students performed statistically lower than Caucasian students; Asian students performed statistically higher than African American students and Hispanic students.

Statistically significant differences in oral reading fluency were found based on educational classifications. Students receiving special education services performed statistically lower than other students, $t(253) = 12.89$, $p < .05$. English language learners also performed statistically lower than other students, $t(25) = 3.4$, $p < .05$.

Reading Comprehension

To analyze the reading comprehension results, 1117 student records were examined. With the range of possible scores from 0-9, the mean score was 7.76, with a standard deviation of 1.47. Descriptive statistics for reading comprehension by ethnicity and educational classification are presented in Table 2.

Table 2

Descriptive statistics for reading comprehension by gender, ethnicity, and educational classification.

Group	*n*	Mean	Standard deviation
	Gender		
Male	547	7.70	1.50
Female	570	7.82	1.43
	Ethnicity		
Caucasian	764	7.89	1.31
Hispanic	75	7.25	1.90
Asian	55	7.53	2.09
African American	35	7.74	1.04
Native American	27	6.67	1.59
Other	103	7.64	1.44
	Educational Classification		
General education	701	7.93	1.30
Special education	176	6.99	1.71
English-language learner	20	5.75	2.51

No statistically significant difference was observed for gender, $t(1115) = -1.4$, $p > .05$. Statistically significant differences, however, were observed for ethnicity, $F(5,1053) = 4.11$, $p < .05$. Levene's test for homogeneity of variances was statistically significant, therefore post hoc analyses were conducted using Bonferroni adjustments to identify differences between groups. Hispanic students performed statistically lower than Caucasian students.

Statistically significant differences in reading comprehension were found based on educational classifications. Students receiving special education services performed statistically lower than other students, $t(219) = 6.65$, $p < .05$. English language learners also performed statistically lower than other students, $t(19) = 3.6$, $p < .05$.

Vocabulary

Descriptive Statistics. The total number of students with valid scores on the vocabulary test was 1127. Given a range of possible scores from 0-25, the mean score for vocabulary was 22.6 words correctly defined, and the standard deviation was 3.44. Descriptive statistics for vocabulary by ethnicity and educational classification are presented in Table 3.

Table 3

Descriptive statistics for vocabulary by gender, ethnicity, and educational classification.

Group	*n*	Mean	Standard deviation
	Gender		
Male	522	22.69	3.21
Female	549	22.44	3.60
	Ethnicity		
Caucasian	772	23.00	2.82
Hispanic	80	20.50	4.56
Asian	56	21.88	4.95
African American	33	21.15	4.65
Native American	26	21.62	4.56
Other	103	22.60	2.93
	Educational Classification		
General education	715	23.18	2.34
Special education	172	19.56	5.20
English-language learner	24	16.71	5.35

No statistically significant difference was observed for gender, $t(1125) = -1.31$, $p > .05$. Statistically significant differences, however, were observed for ethnicity, $F(5,1064) = 11.39$, $p < .05$. Levene's test for homogeneity of variance was statistically significant; therefore post hoc analyses using Bonferroni adjustments were conducted to determine differences between groups. African Americans performed statistically lower than Caucasian students. Hispanic students performed statistically lower than Caucasian students and students with other ethnicities.

Statistically significant differences in vocabulary were found based on educational classifications. Students receiving special education services performed statistically lower than other students, $t(188) = 8.68$, $p < .05$. English language learners also performed statistically lower than other students, $t(23) = 5.44$, $p < .05$.

The internal consistency of the items ranged from $r = .29 - .53$. Therefore, it can be concluded that there is a strong internal consistency across the items without being too high so that items are measuring the same skills.

Option Analysis. Each response selection was analyzed to determine the frequency of selection. Table 4 identifies the number of times that each option was selected, with the overall percentage in parentheses.

Table 4

Option analysis for vocabulary tasks.

Vocabulary Word	Correct Answer	Frequency Response A (%)	Frequency Response B (%)	Frequency Response C (%)
Branch	B	18(2)	1128(97)	17(2)
Calfskin	B	65(6)	1023(88)	70(6)
Cartoon	A	1088(94)	23(2)	50(4)
Chilly	C	39(3)	93(8)	1030(89)
Choose	C	53(5)	23(2)	1083(93)
Club	B	55(5)	837(72)	268(23)
Escape	A	1093(94)	38(3)	29(3)
Explain	A	1020(88)	93(8)	47(4)
False	C	33(3)	44(4)	1086(94)
Fluffy	C	48(4)	38(3)	1076(93)

Green	A	799(69)	298(26)	61(5)
Hike	C	15(1)	16(1)	1132(97)
Lullaby	C	49(4)	34(3)	1074(92)
Lumber	B	70(6)	1006(87)	82(7)
Magnet	B	36(3)	1050(90)	72(6)
Mash	B	28(2)	1113(96)	22(2)
Measure	B	53(5)	1075(93)	33(3)
Mermaid	C	38(3)	40(3)	1081(93)
Merry	C	171(15)	64(6)	924(80)
Often	B	63(5)	976(84)	123(11)
Pitch	A	1117(96)	27(2)	16(1)
Simple	C	27(2)	30(3)	1105(95)
Tornado	B	43(4)	1075(93)	40(3)
Weak	C	26(2)	46(4)	1090(94)
Wink	C	22(2)	18(2)	1120(96)

Item Difficulty. Two statistical procedures were used to determine item difficulty. First, the percentage of students earning a correct score on each word was calculated to determine the overall difficulty. Next, Item Response Theory (IRT) was used to determine the spread of difficulty represented by the set of vocabulary words. IRT modeling software's scale items along a continuum of difficulty levels ranging from the easiest items at the low end of the scale (-3.0) to the most difficulty items at the high end of the scale (+3.0). The results from these analyses are located in Table 5 in order of difficulty, with the most difficult items at the top of the list.

Table 5

Item difficulty for vocabulary.

Word Vocabulary	Percentage of Students with Correct Response	IRT Difficulty Scale
Green	68.7	2.39
Club	72.0	2.13
Merry	79.4	1.49
Often	83.9	1.05
Lumber	86.5	0.71
Explain	87.7	0.59
Calfskin	88.0	0.53
Chilly	87.0	0.49
Magnet	90.3	0.19
Fluffy	92.5	-0.15
Measure	92.4	-0.15
Lullaby	92.3	-0.19
Tornado	92.4	-0.19
Mermaid	92.9	-0.29
False	93.4	-0.31
Choose	93.1	-0.32
Cartoon	94.0	-0.38
Weak	93.7	-0.40
Escape	94.0	-0.50
Simple	95.0	-0.71
Mash	95.7	-0.89

Pitch	96.0	-1.08
Wink	96.3	-1.20
Branch	97.0	-1.33
Hike	97.3	-1.48

Item discrimination. Each item was examined for differential functioning based on gender and educational classification using *t* tests and ethnicity using analysis of variance. No significant differences in item functioning were observed for gender. There are statistically significant differences in item functioning based on ethnicity for the vocabulary words CLUB ($F(5,1061) = 2.66$, $p < .05$) and FALSE ($F(5,1064) = 2.55$, $p < .05$). For CLUB, there appears to be a group effect. Post hoc analyses using Bonferroni adjustments identified that for FALSE, the difference appears to exist between Asian students and Caucasian students. Asian students performed statistically significantly higher than Caucasian students

For students receiving special education services, the vocabulary word CLUB was significantly more difficult than for other students, $t(227) = 2.01$, $p < .05$, but significantly higher on the vocabulary word CHILLY, $t(277) = -2.17$, $p < .05$.

For English language learners, the vocabulary word MERRY ($t(24) = 2.11$, $p < .05$) was statistically significantly more difficult. The vocabulary words FLUFFY ($t(1102) = -9.47$, $p < .05$) and WINK ($t(1099) = -6.44$, $p < .05$) were statistically significantly more easy.

Research Analyses

Correlations

Correlations for the following measures use a significance level of $p < .0083$ to adjust for the combined measures.

Table 6

Correlations for Measures.

Pair of Measures	Correlation	*p*
ORF – Vocabulary	.60	$p < .0083$
ORF – Comprehension	.29	$p < .0083$
ORF – Composite State Assessment in Reading	.41	$p < .0083$
Vocabulary – Comprehension	.37	$p < .0083$
Vocabulary – Composite State Assessment in Reading	.47	$p < .0083$
Comprehension – Composite State Assessment in Reading	.22	$p < .0083$

The correlation between oral reading fluency and vocabulary indicates a strong, positive relationship. However, the association between oral reading fluency and the other measures used in this study is weak to moderate. A moderately strong relationship was found between vocabulary and composite state assessment score for reading, with only weak to moderate correlations between vocabulary and comprehension. A weak to moderate relationship exists between comprehension and composite state assessment score for reading.

Multiple Regression

Multiple regression was calculated to predict how much of the variance in the composite score for the state assessment in reading is explained by each dependent measure. The linear combination of ORF, comprehension, and vocabulary was significantly related to the composite

score for the state assessment in reading, $F(3,1033) = 113.40$, $p < .05$. Approximately 25% of the variance in the composite scores can be explained by the linear combination of ORF, comprehension, and vocabulary. Table 6 summarizes the regression data for the statewide assessment composite score for reading.

Table 7

Regression Summary for the Composite State Score in Reading.

	Unstandardized Coefficients		Standardized Coefficients		95% Confidence Interval for B	
Independent Variables	B	Std. Error	Beta	t	Lower Bound	Upper Bound
ORF	0.09	0.02	.19	5.65*	0.06	0.13
Reading Comprehension	0.66	0.44	.04	1.48	-0.21	1.52
Vocabulary	2.18	0.22	.34	9.78*	1.75	2.62
Constant	150.93	4.56		33.07*	141.98	156.89

* $p < .01$

Discussion

Statistically significant differences were observed for gender, ethnicity, and educational classification for oral reading fluency. Possible explanations for these findings are differences in maturation between genders, differences in instruction based on services provided, and/or discriminating features within the passage. To explain these finding instructional practices would need to be examined for consistency, reading interventions for students with reading difficulties would need to be implemented with fidelity, and/or multiple ORF passages would need to be pilot tested and examined for different groups of students.

The measures of reading comprehension performed differently for Hispanic students, students receiving special education services, and English language learners. In general, minimal variance was found across scores with a mean of 7.76, standard deviation 1.47 out of 9 possible points. This may indicate that the comprehension questions are undemanding for many of the students. More information about student ability may be gained by including additional questions that probe inferential or evaluative comprehension skills.

Overall, the vocabulary measure also was not challenging for most students. The mean score out of 25 points possible was 22.6, with a standard deviation of 3.44. A majority of the items were functioning below average ability level. Only 9 of 25 items were targeting ability levels above the average. Including more challenging items may provide additional information about student knowledge and ability.

Significant differences were observed on the vocabulary measure based on ethnicity and educational classification. Additionally, several items functioned differently based on each population. Possible explanations include differential background knowledge, instruction, and

language barriers. Pilot testing additional words to examine item functioning based on ethnicity and educational classification may explicate these findings.

The correlations between the dependent measures indicate a moderate to strong relationship between each measure. ORF and vocabulary were most strongly correlated. The relationship between the state assessment composite score in reading was strongest between ORF and vocabulary. This also was evidenced in the multiple regression analysis with ORF, comprehension, and vocabulary predicting 25% of the variance in the composite score. ORF and vocabulary scores accounted for the majority of this variance.