An Examination of Test-Retest, Alternate Form Reliability, and

Generalizability Theory Study of the easyCBM Word and Passage

Reading Fluency Assessments:

Grade 3

Bitnara Jasmine Park

Daniel Anderson

Julie Alonzo

Cheng-Fei Lai

Gerald Tindal

University of Oregon

Abstract

This technical report is one in a series of five describing the reliability (test/retest and alternate form) and G–Theory / D–Study research on the easyCBM reading measures, grades 1–5. Data were gathered in the spring of 2011 from a convenience sample of students nested within classrooms at a medium–sized school district in the Pacific Northwest. Due to the length of the results, we present results of each grade level's analysis in its own technical report, sharing a common abstract, introduction, and methods section, while differing in the results and conclusions.

An Examination of Test-Retest, Alternate Form Reliability, and Generalizability Theory

Study of the easyCBM Word and Passage Reading Fluency Assessments: Grade 3

Progress monitoring assessments are a key component of many school improvement efforts, including the Response to Intervention (RTI) approach to meeting students' academic needs. In an RTI approach, teachers first administer a screening or benchmarking assessment to identify students who need supplemental interventions to meet grade-level expectations, then use a series of progress monitoring measures to evaluate the effectiveness of the interventions they are using with the students. When students fail to show expected levels of progress (as indicated by "flat line" scores or little improvement on repeated measures over time), teachers use this information to help them make instructional modifications with the goal of finding an intervention or combination of instructional approaches that will enable each student to make adequate progress toward achieving grade-level proficiency on content standards. In such a system, it is critical to have reliable measures that assess the target construct and are sensitive enough to detect improvement in skill over short periods of time.

Conceptual Framework: Curriculum-Based Measurement and Progress Monitoring

Curriculum-based measurement (CBM), long a bastion of special education, is gaining support among general education teachers seeking a way to monitor the progress their students are making toward achieving grade-level proficiency in key skill and content areas. By definition, CBM is a formative assessment approach. By sampling skills related to the curricular content covered in a given year of instruction yet not specifically associated with a particular textbook, CBMs provide teachers with a snapshot of their students' current level of proficiency in a particular content area as well as a mechanism for tracking the progress students make in gaining desired academic skills throughout the year. Historically, CBMs have been very brief

individually administered measures (Deno, 2003; Good, Gruba, & Kaminski, 2002), yet they are not limited to the one minute timed probes with which many people associate them.

In one of the early definitions of CBM, Deno (1987) stated that "the term curriculum-based assessment, generally refers to any approach that uses direct observation and recording of a student's performance in the local school curriculum as a basis for gathering information to make instructional decisions…The term curriculum-based measurement refers to a specific set of procedures created through a research and development program … and grew out of the *Data-Based Program Modification* system developed by Deno and Mirkin (1977)" (p. 41). He noted that CBM is distinct from many teacher-made classroom assessments in two important respects: (a) the procedures reflect technically-adequate measures ("they possess reliability and validity to a degree that equals or exceeds that of most achievement tests" (p. 41), and (b) "growth is described by an increasing score on a standard, or constant task. The most common application of CBM requires that a student's performance in each curriculum area be measured on a single global task repeatedly across time" (p. 41).

In the three decades since Deno and his colleagues introduced CBM, *progress monitoring probes* as they have come to be called, have increased in popularity, and they are now a regular part of many schools' educational programs (Alonzo, Tindal, & Ketterlin-Geller, & 2006). However, CBMs – even those widely used across the United States – often lack the psychometric properties expected of modern technically-adequate assessments. Although the precision of instrument development has advanced tremendously in the past 30 years with the advent of more sophisticated statistical techniques for analyzing tests on an item by item basis rather than relying exclusively on comparisons of means and standard deviations to evaluate comparability of alternate forms, the world of CBMs has not always kept pace with these statistical advances.

A key feature of assessments designed for progress monitoring is that alternate forms must be as equivalent as possible to allow meaningful interpretation of student performance data across time. Without such cross-form equivalence, changes in scores from one testing occasion to the next are difficult to attribute to changes in student skill or knowledge. Improvements in student scores may, in fact, be an artifact of the second form of the assessment being easier than the form that was administered first. The advent of more sophisticated data analysis techniques (such as the Rasch modeling used in the development of the easyCBM progress monitoring and benchmarking assessments) has made it possible to increase the precision with which we develop and evaluate the quality of assessment tools.

In this technical report, we provide the results of a series of studies to evaluate the technical adequacy of the easyCBM progress monitoring assessments in reading, designed for use with students in Grades 1 - 5. This assessment system was developed to be used by educators interested in monitoring the progress their students make in acquiring skills in the constructs of early literacy (phonemic awareness, phonics), and both word and passage reading fluency. Specifically, we conducted traditional test-retest and alternate form reliability analyses of the easyCBM reading measures. In addition to these more traditional analyses, we applied generalizability theory – a more modern approach to reliability that parses out sources of error variance. As part of the methods section, we briefly outline the purpose and application of generalizability theory.

The easyCBM™ Progress Monitoring Assessments

The online easyCBM™ progress monitoring assessment system, launched in September 2006 as part of a Model Demonstration Center on Progress Monitoring, was initially funded by the Office of Special Education Programs (OSEP). At the time this technical report was

published, there were 92,925 teachers with easyCBM accounts, representing schools and districts spread across every state in the country. During the 2010-2011 school year, the system had an average of 1200 new accounts registered each week, and the popularity of the system continues to grow. In the month of November 2011, alone, 5945 new teachers registered for accounts, with almost 2 million students active on the system at the end of December 2011. The online assessment system provides both universal screener assessments for fall, winter, and spring administration and multiple alternate forms of a variety of progress monitoring measures designed for use in K-8 school settings.

As part of state funding for Response to Intervention (RTI), states need technically-adequate measures for monitoring progress. Given the increasing popularity of the easyCBM online assessment system, it is imperative that a thorough analysis of the measures' technical adequacy be conducted and the results shared with research and practitioner communities. This technical report addresses that need directly, providing the results of a series of studies examining the technical adequacy of the 2009 / 2010 version of the individually-administered easyCBM assessments in reading.

Methods

Data for these analyses were gathered in the spring of 2011 from a convenience sample of students in a mid-sized school district in the Pacific Northwest. Teams of trained research assistants from the University of Oregon administered a battery of easyCBM assessments to students in participating classrooms. Data were gathered in two separate sessions, one week apart. Each day, students were administered a series of alternate forms of grade-appropriate easyCBM assessments in one-on-one settings. Assessors followed standardized administration protocols for all assessments. The assessments were counter-balanced to enable examination of

order effect as well as alternate form reliability, with selected forms repeated across testing

sessions, to allow for test-retest analyses. All assessments were administered in the order

displayed in Appendix A.

Test-Retest and Alternate Form Reliability

We used bivariate correlations to calculate the test-retest and alternate form reliability of

the measures included in this study. These analyses were completed, in part, as a requisite step to

the generalizability theory (G-Theory) analyses. That is, the G-Theory analyses treated each

form as a random observation from the universe of possible forms. The G-Theory analyses thus

assume form equivalence during the d-study prophecy estimations (i.e., the model assumes each

form contributes an equal amount to the measurement process, and that any successive forms

will likewise contribute an equal amount). The comparability of forms had to first be established

to ensure there were no egregious departures.

Generalizability Theory

For our generalizability theory study (G-Study) we calculated the variances associated

persons and two facets: forms and occasions. We then conducted decision studies (D-Studies) to

help determine the necessary conditions for reliable measurement. In this section we first provide

an overview of G- and D-Studies for the two-facet design for readers who may be unfamiliar

with the technique. Readers familiar with G-Theory may want to skip this section and proceed to

the *G-Theory analyses* section.

G-Theory overview. G-theory designs can be crossed or nested. A crossed design is one

that includes students being administered *the same test forms* on both occasions, while a nested

design includes students being administered *different test forms* on both occasions. G-studies are

usually followed up with decision studies (D-study analyses), which provide the number of

levels needed to obtain adequate measurement for each facet. For example, to obtain reliable

estimates of students' ability, should students be administered 1, 2, 3, 4, or 5 forms during any

one occasion? Similarly, does increasing the number of occasions increase the reliability of the

estimate, and at what point is a reliable estimate obtained? The results of the G-study are

analogous to an analysis of variance (ANOVA), while the results of the D-study are similar to a

Spearman-Brown prophecy analysis. Ideally, most of the variance in the G-theory analysis would

be associated with persons, and administering students one test form on one occasion would

result in sufficiently reliable estimates for the D-study.

Absolute and relative error variances are produced during the D-study. The absolute error

variance is the sum of all variance components minus the variance uniquely associated with

persons. That is

$$\sigma_\Delta^2 = \frac{\sigma_F^2}{n_F'} + \frac{\sigma_O^2}{n_O'} + \frac{\sigma_{pF}^2}{n_p' n_F'} + \frac{\sigma_{pO}^2}{n_p' n_O'} + \frac{\sigma_{FO}^2}{n_F' n_O'} + \frac{\sigma_{pFO}^2}{n_p' n_F' n_O'} \qquad (1)$$

where σ_Δ^2 = absolute error variance,

σ_F^2 = variance associated with forms,

σ_O^2 = variance associated with occasions,

σ_{pF}^2 = variance associated with the interaction between persons and forms,

σ_{pO}^2 = variance associated with the interaction between persons and occasions,

σ_{FO}^2 = variance associated with the interaction between forms and occasions,

σ_{pFO}^2 = variance associated with the interaction between persons, forms, and occasions, and

all *n*'s represent the number of factors contributing to the variance component. The single

quotation mark on each *n* represents a value that can be changed to obtain estimates of the

variance with different numbers contributing to the variance estimate – for example, increasing

the number of test forms or testing occasions. Each of these variance components is produced

from the G-study and is reported for the observed n's. The final variance term (person by form by occasion interaction) is generally interpreted as the residual.

The square root of the absolute variances can be interpreted as the "absolute" standard error of measurement (SEM). Absolute variances are generally used to make criterion/domain-referenced decisions (Shavelson & Webb, 2006), or within-student decisions (Hintze, Owen, Shapiro, & Daly, 2000). Relative error variances are used to make normative decisions (i.e., relative to the other persons tested, what is the standard error?). According to Brennan (2001), the square root of the relative error variances can be interpreted essentially identically to the SEM in classical test theory. The relative error variances will nearly always be lower than the absolute variance because only variance components including persons are included. For the two-facet design the relative error variance is defined as

$$\sigma_\delta^2 = \frac{\sigma_{pF}^2}{n_F'} + \frac{\sigma_{pO}^2}{n_O'} + \frac{\sigma_{pFO}^2}{n_F' n_O'} \tag{2}$$

where σ_δ^2 = relative error variance, and all other terms are defined as above. In this paper, we present both the variances and their corresponding square root, which places the value back onto the scale of the measure. For ease of interpretation, we call the square root of the variances the absolute or relative standard error of the measures. Although the analogy is not direct, the interpretation is similar enough that these terms can be used to facilitate understanding. Just as with classical test theory, the SEMs can be used to construct confidence intervals, as in

$$95\% \text{ CI} = X_{pFO} \pm 1.96(\text{SEM}) \tag{3}$$

where X_{pFO} is the score X for person p on form F on occasion O. One of the added benefits of G-theory is the potential to construct both absolute and relative confidence intervals depending on the decision to be made.

Two types of coefficients are generally produced during the D-study analyses: Generalizability or G-coefficients (Ep^2), which are analogous to coefficient alpha in classical test theory (Brennan, 2001) and phi coefficients (Φ), which are an index of the dependability of the measurement process. Just as with the variance components, these two coefficients correspond to absolute (phi) and relative (g) decisions. The phi index of dependability for absolute decisions is given by

$$\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\Delta^2} \tag{4}$$

where all terms are defined as above. In contrast, the g-coefficient for relative decisions is given by

$$Ep^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\delta^2} \tag{5}$$

where all terms are defined as above. Note that the only difference between equations 4 and 5 is the variance component in the denominator, with the phi-coefficient using the absolute error variance term and the g-coefficient using the relative error variance term.

For each analysis, plots can be produced detailing the change in Ep^2 or Φ with increasing the number of testing occasions and forms administered within each occasion. These are generally displayed as line graphs, with each line representing a different *n'* of Facet 1 and the x-axis representing a different *n'* for Facet 2. The plot is simply a visual depiction of the change in reliability coefficients with a corresponding change in the measurement process.

In sum, the G-study provides further information on the sources of error in the measurement process while the D-study provides further information on potential ways that the measurement process could become more dependable. The coefficients to be interpreted depend

upon the use of the measurement tool. If decisions are being made relative to other students (e.g., benchmarking assessments), then the relative error variances and g-coefficients should be interpreted. In contrast, if within-student decisions are being made (e.g., progress-monitoring assessments) then the absolute variances and phi-coefficients should be interpreted.

G-Theory analyses. For this study, all analyses were restricted to groups where a fully crossed design was possible (i.e., all students in the analysis were included in both testing occasions and administered the same test forms). The test forms were often administered in a different order on the separate occasions to mitigate order effects. The forms themselves remained constant across occasions in all analyses. We conducted three G-theory analyses for each of the word reading fluency (WRF) measure types and two analyses for each of the passage reading fluency (PRF) measures. As the table in Appendix A indicates, data from teacher 9 were missing for Occasion 1 across all measure types. Teacher 9 was thus dropped from all analyses. All data were examined in a fully-crossed two-facet design. The first facet in the analysis, *form*, was generally counter-balanced across occasions. The second facet was *occasion*.

For the first WRF analysis, data for Teacher 10 were analyzed and test forms 14 and 16 were examined in a non-counterbalanced design. The second analysis examined data from forms 14 and 15 for teacher 11 in a counterbalanced design, while the third examined data from forms 11, 12, and 13 for teacher 12 in a partially counterbalanced design (see Table 2 in Appendix A). For the first PRF analysis, data were collapsed for Teacher 10 and 11 to examine the generalizability of forms 14, 15, and 16 in a partially counterbalanced design. Forms 11, 12 and 13 were examined for teacher 12 in the second analysis in a partially counterbalanced design.

For all G-theory analyses, forms were analyzed in ascending order regardless of administration order. For example, for the first analysis for WRF, the order of administration for

forms 14 and 16 varied by the occasion. However, during the analysis the data were analyzed for forms 14 and 16 on the first occasion and forms 14 and 16 on the second occasion. In other words, the analysis did not attempt to replicate the administration order because the counterbalanced design was intended to mitigate any order effects. All G-theory analyses were conducted using the SPSS macro produced by Mushquash and O'Connor (2006).

In our results section, we present the results of our G-Studies through an analysis of variance (ANOVA) table detailing the variance associated with each facet of the measurement process as well as all interactions among facets. We then present the error variances and G-coefficients for the design used before presenting the D-Study prophecy estimations results. The D-Study error variance estimates are also presented in their standard error form (i.e., $\sqrt{\sigma^2(\Delta_p)}$ and $\sqrt{\sigma^2(\delta_p)}$ for absolute and relative standard errors respectively), which places the error term back on the scale of the measure and can be used to construct confidence intervals for any individual student's score for any of the measurement designs investigated. Following the error variance estimates, the prophesized G- and Phi-coefficient estimates are presented. Finally a plot was produced for each analysis detailing the estimated change in Ep^2 (labeled on the y-axis as "Mean gstat") with increasing the number of testing occasions and forms administered within each occasion. Each line on the graph represents a different number of testing occasions, ranging from 1-5, while the x-axis represents the number of forms within any occasion. The plot is simply a visual depiction of the G-coefficients table for the corresponding analysis.

Results

The results of the grade 3 reading assessments are presented below, organized by type of measure.

Word Reading Fluency

Descriptive statistics are presented in Tables 1 and 2. Test-retest reliability results are presented in Table 3. Correlations between each of the six Word Reading Passage forms are presented in Table 4.

Table 1
Descriptive Statistics for Grade 3 Word Reading Fluency Measures: Session 1

Test Form	*n*	Min	Max	*M*	*SD*
WRF3.11.1	17	28	86	60.88	16.49
WRF3.12.1	17	33	87	58.24	16.37
WRF3.13.1	17	31	87	58.18	16.55
WRF3.14.1	31	24	100	69.61	17.21
WRF3.15.1	31	25	100	73.61	16.84
WRF3.16.1	31	25	106	68.42	17.08

Table 2
Descriptive Statistics for Grade 3 Word Reading Fluency Measures: Session 2

Test Form	*n*	Min	Max	*M*	*SD*
WRF3.11.2	36	23	123	69.64	21.36
WRF3.12.2	36	19	107	68.67	19.92
WRF3.13.2	36	14	117	69.61	21.91
WRF3.14.2	53	17	107	67.42	18.60
WRF3.15.2	25	19	124	64.76	22.29
WRF3.16.2	28	25	96	71.29	17.39

Test-Retest Reliability. To evaluate test-retest reliability, we correlated performance on each form of the WRF measure that was administered across the two testing sessions. Table 3 present results of these analyses. Overall, test-retest reliability was moderately strong, ranging from .67 to .92.

Table 3
Test-retest Reliability Results: Word Reading Fluency

Test Form	WRF3.11.2	WRF3.12.2	WRF3.13.2	WRF3.14.2	WRF3.15.2	WRF3.16.2
WRF3.11.1	0.92					
WRF3.12.1		0.90				
WRF3.13.1			0.83			
WRF3.14.1				0.74		
WRF3.15.1					0.67	
WRF3.16.1						0.69

Alternate Form Reliability. Alternate form reliability was analyzed using bi-variate correlations. We present the correlations between the different forms of each WRF measure in Table 4. We found a moderate positive relationship between the alternate forms, with correlations ranging from .72 to .92.

Table 4
Correlation between Alternate Forms of Grade 3 Word Reading Fluency Measure

Test Form	WRF3.12.1	WRF3.13.1	WRF3.15.2	WRF3.16.2
WRF3.11.1	0.92	0.87		
WRF3.12.1		0.84		
WRF3.14.2			0.87	0.72
WRF3.15.2				0.81

G-study / D-study results. The results of the test-retest and alternate-form reliability analyses suggested acceptable form equivalence for subsequent G-Theory analyses. For the three Word Reading Fluency analyses, 48%, 70%, and 77% of the variance was associated with the 20, 9, and 17 persons included in the analysis, 0% was associated with forms, and 0% was associated with occasion. The relative error variance was 35.08, 49.34, and 12.57 for the first, second, and third analysis, respectively. The absolute variance was 42.51, 53.03, and 26.14, respectively. The G-Coefficients were .74 for the first analysis, .85 for the second analysis, and .95 for the third analysis, while the phi coefficients were .70, .84 and .91, respectively.

	Word Reading Fluency: Forms 14 & 16 (teacher 10)	

Grade 3 WRF: Forms 14 & 16

Generalizability ANOVA Table

Facet	*df*	SS	MS	Variance	Proportion
Persons	19	10133.2	533.326	98.255	0.484
Forms	1	110.45	110.45	0.000	0.000
Occasions	1	672.8	672.8	9.279	0.046
Person*Forms	19	1723.55	90.713	31.121	0.153
Person*Occasion	19	1483.2	78.063	24.796	0.122
Forms*Occasion	1	252.05	252.05	11.179	0.055
Person*Forms*Occasions (Residual)	19	540.95	28.471	28.471	0.14

Note. Analysis included 20 students, with 2 forms (14 & 16) on 2 occasions.

Error Variances:

Relative, $\sigma^2(\delta_p)$ | Absolute, $\sigma^2(\Delta_p)$
  35.076   42.511

G-coefficients:

  G: Ep^2 | Phi: Φ
   .737   .698

Grade 3 WRF: Forms 14 & 16

D-Study Absolute Error Variances, $\sigma^2(\Delta_p)$

n forms	n occasions				
	1	2	3	4	5
1	104.846	67.984	55.696	49.552	45.866
2	69.461	**42.511**	33.527	29.036	26.341
3	57.665	34.020	26.138	22.197	19.832
4	51.768	29.774	22.443	18.777	16.578
5	48.229	27.227	20.226	16.725	14.625

Grade 3 WRF: Forms 14 & 16

D-Study Absolute Standard Errors, $\sigma(\Delta_p)$

n forms	n occasions				
	1	2	3	4	5
1	10.239	8.245	7.463	7.039	6.772
2	8.334	**6.520**	5.790	5.389	5.132
3	7.594	5.833	5.113	4.711	4.453
4	7.195	5.457	4.737	4.333	4.072
5	6.945	5.218	4.497	4.090	3.824

Grade 3 WRF: Forms 14 & 16

D-Study Relative Error Variances, $\sigma^2(\delta_p)$

n forms	n occasions				
	1	2	3	4	5
1	84.388	57.755	48.877	44.438	41.774
2	54.592	**35.076**	28.571	25.318	23.367
3	44.660	27.517	21.802	18.945	17.231
4	39.694	23.737	18.418	15.759	14.163
5	36.714	21.469	16.388	13.847	12.322

Grade 3 WRF: Forms 14 & 16

D-Study Relative Standard Errors, $\sigma(\delta_p)$

n forms	n occasions				
	1	2	3	4	5
1	9.186	7.600	6.991	6.666	6.463
2	7.389	**5.922**	5.345	5.032	4.834
3	6.683	5.246	4.669	4.353	4.151
4	6.300	4.872	4.292	3.970	3.763
5	6.059	4.633	4.048	3.721	3.510

Grade 3 WRF: Forms 14 & 16

D-Study G Coefficients, Ep^2

n forms	*n* occasions				
	1	2	3	4	5
1	0.538	0.630	0.668	0.689	0.702
2	0.643	**0.737**	0.775	0.795	0.808
3	0.688	0.781	0.818	0.838	0.851
4	0.712	0.805	0.842	0.862	0.874
5	0.728	0.821	0.857	0.876	0.889

Grade 3 WRF: Forms 14 & 16

D-Study Phi Coefficients, Φ

n forms	*n* occasions				
	1	2	3	4	5
1	0.484	0.591	0.638	0.665	0.682
2	0.586	**0.698**	0.746	0.772	0.789
3	0.630	0.743	0.790	0.816	0.832
4	0.655	0.767	0.814	0.840	0.856
5	0.671	0.783	0.829	0.855	0.870

G-Coefficient
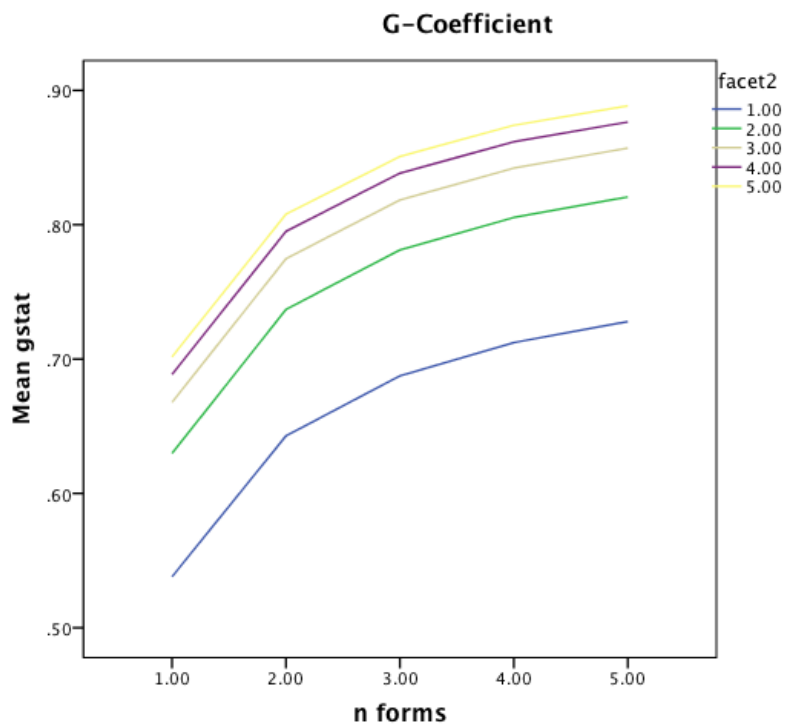
Word Reading Fluency: Forms 14 & 15 (teacher 11)

Grade 3 WRF: Forms 14 & 15

Generalizability ANOVA Table

Facet	*df*	SS	MS	Variance	Proportion
Persons	8	10372.06	1296.507	276.875	0.699
Forms	1	78.028	78.028	4.722	0.012
Occasions	1	220.028	220.028	2.667	0.007
Person*Forms	8	146.722	18.34	0.000	0.000
Person*Occasion	8	1578.722	197.34	85.333	0.215
Forms*Occasion	1	1.361	1.361	0.000	0.000
Person*Forms*Occasions (Residual)	8	213.389	26.674	26.674	0.067

Note. Analysis included 9 students, with 2 forms (14 & 15) on 2 occasions.

Error Variances:

Relative, $\sigma^2(\delta_p)$ | Absolute, $\sigma^2(\Delta_p)$
    49.335   53.030

G-coefficients:

    G: Ep^2 | Phi: Φ
    .849   .839

Grade 3 WRF: Forms 14 & 15

D-Study Absolute Error Variances, $\sigma^2(\Delta_p)$

n forms	*n* occasions				
	1	2	3	4	5
1	119.396	62.059	42.947	33.391	27.657
2	103.698	**53.030**	36.140	27.695	22.628
3	98.465	50.020	33.871	25.797	20.952
4	95.849	48.515	32.737	24.848	20.114
5	94.279	47.612	32.056	24.278	19.611

Grade 3 WRF: Forms 14 & 15

D-Study Absolute Standard Errors, $\sigma(\Delta_p)$

n forms	*n* occasions				
	1	2	3	4	5
1	10.927	7.878	6.553	5.778	5.259
2	10.183	**7.282**	6.012	5.263	4.757
3	9.923	7.072	5.820	5.079	4.577
4	9.790	6.965	5.722	4.985	4.485
5	9.710	6.900	5.662	4.927	4.428

Grade 3 WRF: Forms 14 & 15

D-Study Relative Error Variances, $\sigma^2(\delta_p)$

n forms	*n* occasions				
	1	2	3	4	5
1	112.007	56.003	37.336	28.002	22.401
2	98.670	**49.335**	32.890	24.668	19.734
3	94.225	47.112	31.408	23.556	18.845
4	92.002	46.001	30.667	23.000	18.400
5	90.668	45.334	30.223	22.667	18.134

Grade 3 WRF: Forms 14 & 15

D-Study Relative Standard Errors, $\sigma(\delta_p)$

n forms	*n* occasions				
	1	2	3	4	5
1	10.583	7.484	6.110	5.292	4.733
2	9.933	**7.024**	5.735	4.967	4.442
3	9.707	6.864	5.604	4.853	4.341
4	9.592	6.782	5.538	4.796	4.290
5	9.522	6.733	5.498	4.761	4.258

Grade 3 WRF: Forms 14 & 15

D-Study G Coefficients, Ep^2

n forms	*n* occasions				
	1	2	3	4	5
1	0.712	0.832	0.881	0.908	0.925
2	0.737	**0.849**	0.894	0.918	0.933
3	0.746	0.855	0.898	0.922	0.936
4	0.751	0.858	0.900	0.923	0.938
5	0.753	0.859	0.902	0.924	0.939

Grade 3 WRF: Forms 14 & 15

D-Study Phi Coefficients, Φ

n forms	*n* occasions				
	1	2	3	4	5
1	0.699	0.817	0.866	0.892	0.909
2	0.728	**0.839**	0.885	0.909	0.924
3	0.738	0.847	0.891	0.915	0.930
4	0.743	0.851	0.894	0.918	0.932
5	0.746	0.853	0.896	0.919	0.934

G-Coefficient
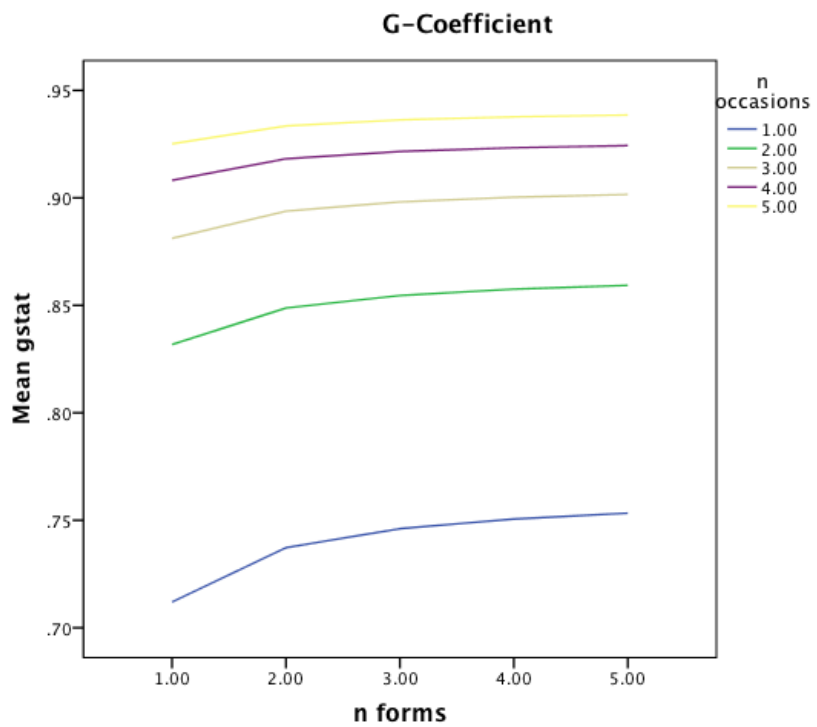
Word Reading Fluency: Forms 11, 12, & 13 (teacher 12)

Grade 3 WRF: Forms 11, 12 & 13

Generalizability ANOVA Table

Facet	*df*	SS	MS	Variance	Proportion
Persons	16	25577.31	1598.582	253.863	0.768
Forms	2	23.608	11.804	0.000	0.000
Occasions	1	1468.324	1468.324	27.662	0.084
Person*Forms	32	1599.392	49.981	9.614	0.029
Person*Occasion	16	898.843	56.178	8.475	0.026
Forms*Occasion	2	64.235	32.118	0.08	0.000
Person*Forms*Occasions (Residual)	32	984.098	30.753	30.753	0.093

Note. Analysis included 17 students, with 3 forms (11, 12 & 13) on 2 occasions.

Error Variances:

Relative, $\sigma^2(\delta_p)$ | Absolute, $\sigma^2(\Delta_p)$
  12.568   26.412

G-coefficients:

  G: Ep^2 | Phi: Φ
   .953   .906

Grade 3 WRF: Forms 11, 12 & 13

D-Study Absolute Error Variances, $\sigma^2(\Delta_p)$

n forms	*n* occasions				
	1	2	3	4	5
1	76.585	43.099	31.938	26.357	23.008
2	56.361	30.584	21.992	17.695	15.118
3	49.620	**26.412**	18.676	14.808	12.488
4	46.249	24.326	17.019	13.365	11.173
5	44.227	23.075	16.024	12.499	10.384

Grade 3 WRF: Forms 11, 12 & 13

D-Study Absolute Standard Errors, $\sigma(\Delta_p)$

n forms	*n* occasions				
	1	2	3	4	5
1	8.751	6.565	5.651	5.134	4.797
2	7.507	5.530	4.690	4.207	3.888
3	7.044	**5.139**	4.322	3.848	3.534
4	6.801	4.932	4.125	3.656	3.343
5	6.650	4.804	4.003	3.535	3.222

Grade 3 WRF: Forms 11, 12 & 13

D-Study Relative Error Variances, $\sigma^2(\delta_p)$

n forms	*n* occasions				
	1	2	3	4	5
1	48.842	29.228	22.690	19.421	17.460
2	28.658	16.733	12.757	10.770	9.577
3	21.931	**12.568**	9.447	7.886	6.950
4	18.567	10.485	7.791	6.444	5.636
5	16.548	9.236	6.798	5.579	4.848

Grade 3 WRF: Forms 11, 12 & 13

D-Study Relative Standard Errors, $\sigma(\delta_p)$

n forms	*n* occasions				
	1	2	3	4	5
1	6.989	5.406	4.763	4.407	4.179
2	5.353	4.091	3.572	3.282	3.095
3	4.683	**3.545**	3.074	2.808	2.636
4	4.309	3.238	2.791	2.539	2.374
5	4.068	3.039	2.607	2.362	2.202

Grade 3 WRF: Forms 11, 12 & 13

D-Study G Coefficients, Ep^2

n forms	*n* occasions				
	1	2	3	4	5
1	0.839	0.897	0.918	0.929	0.936
2	0.899	0.938	0.952	0.959	0.964
3	0.920	**0.953**	0.964	0.970	0.973
4	0.932	0.960	0.970	0.975	0.978
5	0.939	0.965	0.974	0.978	0.981

Grade 3 WRF: Forms 11, 12 & 13

D-Study Phi Coefficients, Φ

n forms	*n* occasions				
	1	2	3	4	5
1	0.768	0.855	0.888	0.906	0.917
2	0.818	0.892	0.920	0.935	0.944
3	0.836	**0.906**	0.931	0.945	0.953
4	0.846	0.913	0.937	0.950	0.958
5	0.852	0.917	0.941	0.953	0.961

G-Coefficient
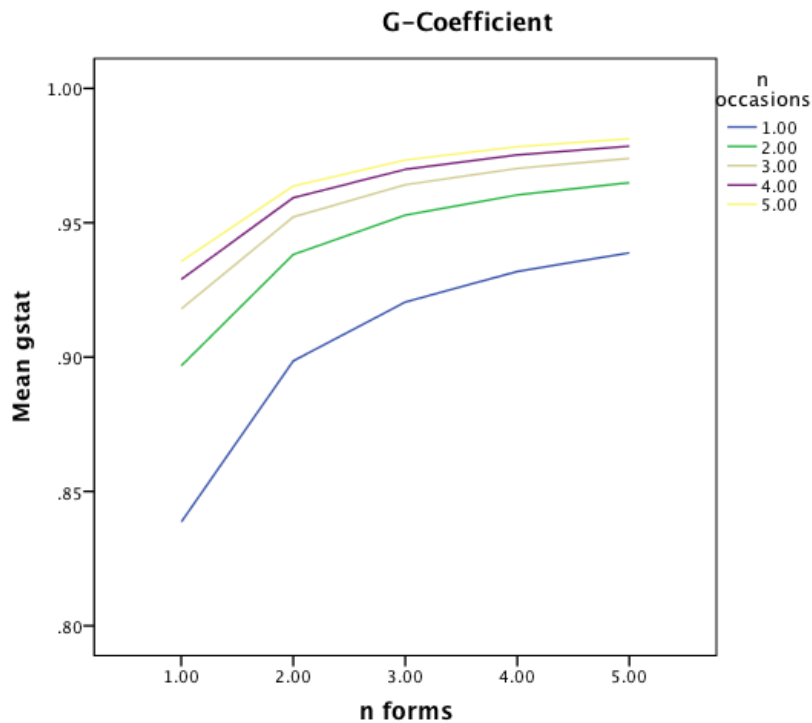
Passage Reading Fluency

Descriptive statistics are presented in Tables 5 and 6. Test-retest reliability results are presented in Table 7. Correlations between each of the six Word Reading Passage forms are presented in Table 8.

Table 5
Descriptive Statistics for Grade 3 Passage Reading Fluency Measures: Session 1

Test Form	*n*	Min	Max	*M*	*SD*
PRF3.11.1	17	51	167	112.35	30.45
PRF3.12.1	17	46	148	109.71	26.06
PRF3.13.1	17	52	161	110.06	29.33
PRF3.14.1	30	45	219	129.77	38.92
PRF3.15.1	31	52	264	137.84	42.88
PRF3.16.1	31	37	227	126.65	35.96

Table 6
Descriptive Statistics for Grade 3 Passage Reading Fluency Measures: Session 2

Test Form	*n*	Min	Max	*M*	*SD*
PRF3.11.2	36	20	244	132.53	44.85
PRF3.12.2	36	24	209	127.72	39.78
PRF3.13.2	36	24	227	126.83	42.34
PRF3.14.2	53	47	257	131.23	40.74
PRF3.15.2	53	51	255	132.94	43.02
PRF3.16.2	53	43	209	122.43	36.92

Test-Retest Reliability. To evaluate test-retest reliability, we correlated performance on each form of the PRF measure that was administered across the two testing sessions. Table 7 present results of these analyses. Overall, test-retest reliability was strong, ranging from .84 to .94.

Table 7
Test-retest Reliability Results: Passage Reading Fluency

Test Form	PRF3.11.2	PRF3.12.2	PRF3.13.2	PRF3.14.2	PRF3.15.2	PRF3.16.2
PRF3.11.1	0.94					
PRF3.12.1		0.87				
PRF3.13.1			0.84			
PRF3.14.1				0.90		
PRF3.15.1					0.90	
PRF3.16.1						0.89

Alternate Form Reliability. Alternate form reliability was analyzed using bi-variate correlations. We present the correlations between the different forms of each PRF measure in Table 8. We found a strong positive relationship between the alternate forms, with correlations ranging from .92 to .96.

Table 8
Correlation between Alternate Forms of Grade 3 Passage Reading Fluency Measure

Test Form	PRF3.12.1	PRF3.13.1	PRF3.15.1	PRF3.16.1
PRF3.11.1	0.94	0.92		
PRF3.12.1		0.95		
PRF3.14.1			0.95	0.96
PRF3.15.1				0.95

G-study / D-study results. The results of the test-retest and alternate-form reliability analyses suggested acceptable form equivalence for subsequent G-Theory analyses. For the two Passage Reading Fluency analyses, 82% and 81% of the variance was associated with the 28 and 17 persons included in the analysis, 0% was associated with forms, and 0% was associated with occasion. The relative error variance was 70.97 and 22.94 for the first and second analysis, respectively. The absolute variance was 97.12 and 61.09, respectively. The G-Coefficients were .95 for the first analysis and .97 for the second analysis, while the phi coefficients were .93for both analyses.

Passage Reading Fluency: Forms 14, 15 & 16 (teachers 10 & 11)

Grade 3 PRF: Forms 14, 15 & 16
Generalizability ANOVA Table

Facet	*df*	SS	MS	Variance	Proportion
Persons	27	211919.5	7848.872	1237.175	0.824
Forms	2	2990.393	1495.196	21.825	0.015
Occasions	1	3483.482	3483.482	36.576	0.024
Person*Forms	54	9417.94	174.406	56.669	0.038
Person*Occasion	27	8437.018	312.482	83.805	0.056
Forms*Occasion	2	319.321	159.661	3.521	0.002
Person*Forms*Occasions (Residual)	54	3297.679	61.068	61.068	0.041

Note. Analysis included 28 students, with 3 forms (14, 15 & 16) on 2 occasions.

Error Variances:
Relative, $\sigma^2(\delta_p)$ | Absolute, $\sigma^2(\Delta_p)$
    70.970   97.120

G-coefficients:

G: Ep^2	Phi: Φ
.946	.927

Grade 3 PRF: Forms 14, 15 & 16

D-Study Absolute Error Variances, $\sigma^2(\Delta_p)$

	n occasions				
n forms	1	2	3	4	5
1	263.464	170.979	140.151	124.737	115.488
2	191.923	115.585	90.139	77.416	69.782
3	168.075	**97.120**	73.468	61.642	54.547
4	156.152	87.888	65.133	53.756	46.929
5	148.998	82.348	60.132	49.024	42.359

Grade 3 PRF: Forms 14, 15 & 16

D-Study Absolute Standard Errors, $\sigma(\Delta_p)$

	n occasions				
n forms	1	2	3	4	5
1	16.232	13.076	11.839	11.169	10.747
2	13.854	10.751	9.494	8.799	8.354
3	12.964	**9.855**	8.571	7.851	7.386
4	12.496	9.375	8.071	7.332	6.850
5	12.206	9.075	7.754	7.002	6.508

Grade 3 PRF: Forms 14, 15 & 16

D-Study Relative Error Variances, $\sigma^2(\delta_p)$

n forms	*n* occasions				
	1	2	3	4	5
1	201.542	129.105	104.960	92.887	85.644
2	142.673	85.504	66.447	56.919	51.202
3	123.050	**70.970**	53.610	44.930	39.722
4	113.239	63.703	47.191	38.935	33.982
5	107.352	59.343	43.340	35.338	30.537

Grade 3 PRF: Forms 14, 15 & 16

D-Study Relative Standard Errors, $\sigma(\delta_p)$

n forms	*n* occasions				
	1	2	3	4	5
1	14.197	11.362	10.245	9.638	9.254
2	11.945	9.247	8.152	7.544	7.156
3	11.093	**8.424**	7.322	6.703	6.303
4	10.641	7.981	6.870	6.240	5.829
5	10.361	7.703	6.583	5.945	5.526

Grade 3 PRF: Forms 14, 15 & 16

D-Study G Coefficients, Ep^2

n forms	*n* occasions				
	1	2	3	4	5
1	0.860	0.906	0.922	0.930	0.935
2	0.897	0.935	0.949	0.956	0.960
3	0.910	**0.946**	0.958	0.965	0.969
4	0.916	0.951	0.963	0.969	0.973
5	0.920	0.954	0.966	0.972	0.976

Grade 3 PRF: Forms 14, 15 & 16

D-Study Phi Coefficients, Φ

n forms	*n* occasions				
	1	2	3	4	5
1	0.824	0.879	0.898	0.908	0.915
2	0.866	0.915	0.932	0.941	0.947
3	0.880	**0.927**	0.944	0.953	0.958
4	0.888	0.934	0.950	0.958	0.963
5	0.893	0.938	0.954	0.962	0.967

G-Coefficient
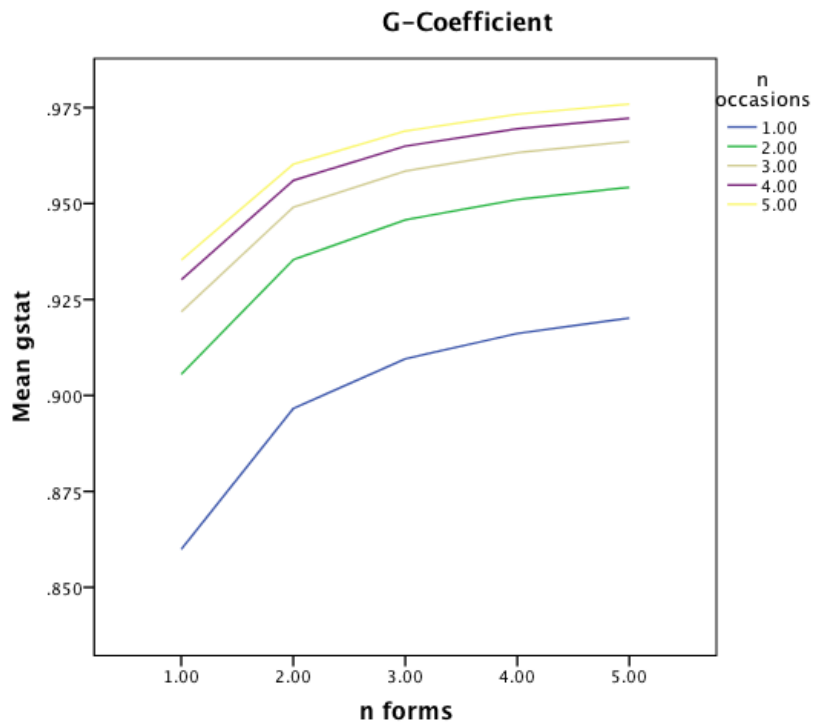
Passage Reading Fluency: Forms 11, 12, & 13 (teacher 12)

Grade 3 PRF: Forms 11, 12 & 13

Generalizability ANOVA Table

Facet	*df*	SS	MS	Variance	Proportion
Persons	16	77777.98	4861.124	789.354	0.809
Forms	2	803.843	401.922	8.008	0.008
Occasions	1	3756.48	3756.48	69.929	0.072
Person*Forms	32	2469.49	77.172	0	0
Person*Occasion	16	2202.02	137.626	15.944	0.016
Forms*Occasion	2	284.549	142.275	3.087	0.003
Person*Forms*Occasions (Residual)	32	2873.451	89.795	89.795	0.092

Note. Analysis included 17 students, with 3 forms (11, 12 & 13) on 2 occasions.

Error Variances:

Relative, $\sigma^2(\delta_p)$ | Absolute, $\sigma^2(\Delta_p)$
       22.938   61.086

G-coefficients:

   G: Ep^2 | Phi: Φ
    .972   .928

Grade 3 PRF: Forms 11, 12 & 13

D-Study Absolute Error Variances, $\sigma^2(\Delta_p)$

n forms	n occasions				
	1	2	3	4	5
1	186.763	97.385	67.593	52.697	43.759
2	136.318	70.161	48.109	37.082	30.467
3	119.503	**61.086**	41.614	31.878	26.036
4	111.095	56.549	38.366	29.275	23.821
5	106.051	53.826	36.418	27.714	22.491

Grade 3 PRF: Forms 11, 12 & 13

D-Study Absolute Standard Errors, $\sigma(\Delta_p)$

n forms	n occasions				
	1	2	3	4	5
1	13.666	9.868	8.221	7.259	6.615
2	11.676	8.376	6.936	6.089	5.520
3	10.932	**7.816**	6.451	5.646	5.103
4	10.540	7.520	6.194	5.411	4.881
5	10.298	7.337	6.035	5.264	4.742

Grade 3 PRF: Forms 11, 12 & 13

D-Study Relative Error Variances, $\sigma^2(\delta_p)$

n forms	*n* occasions				
	1	2	3	4	5
1	105.739	52.869	35.246	26.435	21.148
2	60.841	30.421	20.280	15.210	12.168
3	45.875	**22.938**	15.292	11.469	9.175
4	38.392	19.196	12.797	9.598	7.678
5	33.903	16.951	11.301	8.476	6.781

Grade 3 PRF: Forms 11, 12 & 13

D-Study Relative Standard Errors, $\sigma(\delta_p)$

n forms	*n* occasions				
	1	2	3	4	5
1	10.283	7.271	5.937	5.141	4.599
2	7.800	5.516	4.503	3.900	3.488
3	6.773	**4.789**	3.910	3.387	3.029
4	6.196	4.381	3.577	3.098	2.771
5	5.823	4.117	3.362	2.911	2.604

Grade 3 PRF: Forms 11, 12 & 13

D-Study G Coefficients, Ep^2

n forms	*n* occasions				
	1	2	3	4	5
1	0.882	0.937	0.957	0.968	0.974
2	0.928	0.963	0.975	0.981	0.985
3	0.945	**0.972**	0.981	0.986	0.989
4	0.954	0.976	0.984	0.988	0.990
5	0.959	0.979	0.986	0.989	0.991

Grade 3 PRF: Forms 11, 12 & 13

D-Study Phi Coefficients, Φ

n forms	*n* occasions				
	1	2	3	4	5
1	0.809	0.890	0.921	0.937	0.947
2	0.853	0.918	0.943	0.955	0.963
3	0.869	**0.928**	0.950	0.961	0.968
4	0.877	0.933	0.954	0.964	0.971
5	0.882	0.936	0.956	0.966	0.972

G-Coefficient
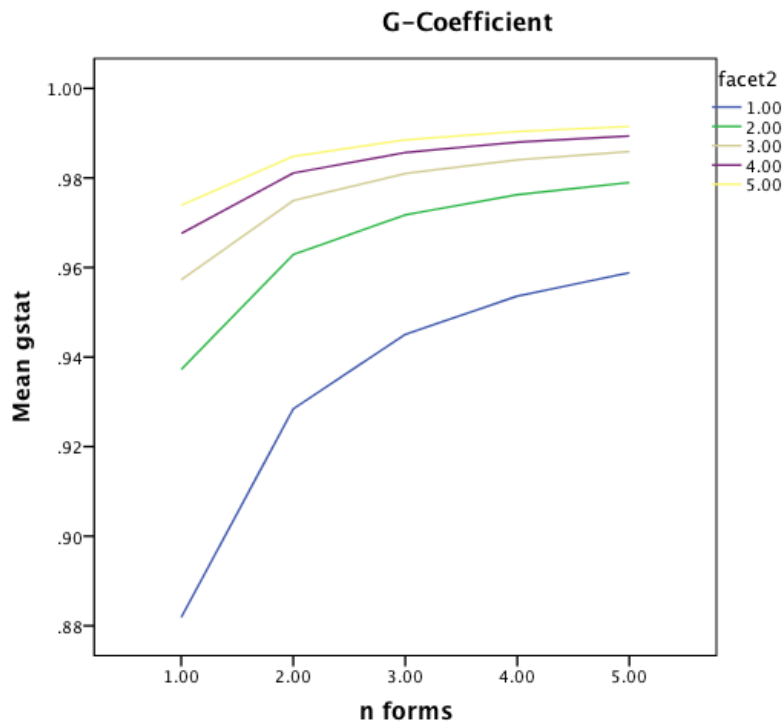
Discussion

The test-retest and alternate form reliability results of this study provide moderate to high evidence of the reliability of the easyCBM grade three WRF and PRF measures, with moderately high test-retest reliability and moderate to high correlations between the alternate forms of the WRF and PRF measures.

The results of the G- and D-studies were less encouraging overall, with 48% - 81% of the total variance associated with persons and the predicted Ep^2 for one test form on one occasion ranging from .54 to .84. The first analysis, which investigated forms 14 & 16 with Teacher 10, was particularly weak with roughly 15% of the total variance attributable to a person by form interaction, 12% to a person by occasion interaction, and 14% to a person by form by occasion interaction. Using .8 as the cutoff for acceptable reliability of relative decisions (Ep^2) the results suggest that students should be tested with at least 4 forms on 2 occasions or 3 forms on 3

occasions. Clearly these results are below optimal levels for practice, suggesting the need for

additional research. The third analysis, which investigated forms 11, 12, and 13 with Teacher 12,

were much better in comparison. Indeed, testing students with one form on one occasion would

be predicted to meet the .8 cutoff for relative decisions. In sum, the results across grade 3 were

inconsistent, with the first two analyses displaying poorer results overall than the third analysis.

We can only speculate on why the results were poorer in these first two analyses, but one

common element was form 14. Although it would be premature to attribute the poor results to

this form, it does warrant further investigation.

References

Alonzo, J., Tindal, G., & Ketterlin-Geller, L.R. (2006). General outcome measures of basic skills in reading and math. In L. Florian (Ed.), Handbook of Special Education. Thousand Oaks, CA: Sage.

Brennan, R. L. (2001). Statistics for social science and public policy: Generalizability theory. New York: Springer.

Deno, S. L. (2003). Developments in curriculum-based measurements. *The Journal of Special Education, 37*, 184-192.

Deno, S. (1987). Curriculum-based measurement. *Teaching Exceptional Children.* (Fall), 41-47.

Deno, S. L., & Mirkin, P. M. (1977). *Data based program modification.* Minneapolis, MN: University of Minnesota Leadership Training Institute/Special Education.

Good, R. H., Gruba, J., & Kaminski, R. A. (2002). Best practices in Using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in an Outcomes-Driven Model. In A. Thomas and J. Grimes (Eds.). *Best Practices in School Psychology IV* (pp.679-700). Washington, DC: National Association of School Psychologists.

Hintze, J. M., Owen, S. V., Shapiro, E. S., and Daly, E. J. (2000). Research design and methodology section: Generalizability of oral reading fluency measures: Application of G theory to curriculum-based measurement. *School Psychology Quarterly, 15*, 52-68.

Mushquash, C., & O'connor, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods, 38*, 542-547.

Shavelson, R. J., & Webb, N. M. (2006). Generalizability theory. In Green, J. L., Camilli, G. & Elmore, P. B. (Eds.), *Complementary Methods for Research in Education,* (pp. 309-322). (3rd ed.) Washington, DC: AERA.

Appendix A

Table 1

Full test form administration order

Teacher	Word Reading Fluency		Passage Reading Fluency	
	Occasion 1	Occasion 2	Occasion 1	Occasion 2
9	-	12 – 13 – 11	-	12 – 13 – 11
10	16 – 15 – 14	16 – 14	16 – 15 – 14	16 – 14 – 15
11	14 – 15 – 16	15 – 14	14 – 15 – 16	15 – 16 – 14
12	13 – 12 – 11	13 – 11 – 12	13 – 12 – 11	13 – 11 – 12

Table 2

Test forms used in G-Theory analyses

Teacher	Word Reading Fluency		Passage Reading Fluency	
	Occasion 1	Occasion 2	Occasion 1	Occasion 2
10	16 – 14	16 – 14	16 – 15 – 14	16 – 14 – 15
11	14 – 15	15 – 14	14 – 15 – 16	15 – 16 – 14
12	13 – 12 – 11	13 – 11 – 12	13 – 12 – 11	13 – 11 – 12