Technical Report # 1217

An Examination of Test-Retest, Alternate Form Reliability, and Generalizability Theory Study of the easyCBM Reading Assessments: Grade 2

Daniel Anderson

Cheg-Fei Lai

Bitnara Jasmine Park

Julie Alonzo

Gerald Tindal

University of Oregon

BRT
behavioral research & teaching

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

Note: Funds for this data set used to generate this report come from a federal grant awarded to the UO from the U.S. Department of Education, Institute for Education Sciences: Reliability and Validity Evidence for Progress Measures in Reading. U.S. Department of Education, Institute for Education Sciences. R324A100014. June 2010 - June 2014. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

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 Reading Assessments: Grade 2

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

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

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

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

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 ($\mathrm{E}p^2$), 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

$$\mathrm{E}p^2 = \frac{\sigma_p^2}{\sigma_p^2+\sigma_\delta^2} \tag{5}$$

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 $\mathrm{E}p^2$ 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. Data for this study were analyzed in a two-facet fully crossed design (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 two G-theory analyses for each of the word reading fluency (WRF) and passage reading fluency (PRF) measure types The first facet in the analysis, *form*, was generally counter-balanced across occasions. The second facet was *occasion*.

For the first WRF analysis, data were collapsed for Teachers 5 and 6 and test forms 11, 12, and 13 were examined. The second analysis was identical but included students instructed by teachers 7 and 8 and test forms 14, 15, and 16 were examined. For the first PRF analysis, data were again collapsed for Teachers 5 and 6 and test forms 11, 12, and 13 were examined. Similarly, data from teachers 7 and 8 were collapsed for the second analysis and test forms 14, 15, and 16 were examined. See Appendix A for the full administration order by teacher.

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 11, 12, and 13 varied by the teacher and occasion. However, during the analysis the data were analyzed for forms 11, 12, and 13 on the first occasion and forms 11, 12, and 13 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 $\mathrm{E}p^2$ (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 2 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 6 forms are presented in Table 4.

Table 1
Descriptive Statistics for Grade 2 Word Reading Fluency Measures: Session 1

Test Form	*N*	Minimum	Maximum	Mean	Std. Deviation
WRF2.11.1	34	18	103	59.94	20.36
WRF2.12.1	34	14	113	61.09	22.12
WRF2.13.1	34	26	96	59.71	19.14
WRF2.14.1	50	11	109	68.12	19.17
WRF2.15.1	50	11	97	68.96	17.64
WRF2.16.1	50	7	100	71.88	20.59

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

Test Form	*N*	Minimum	Maximum	Mean	Std. Deviation
WRF2.11.2	48	13	100	60.08	23.39
WRF2.12.2	48	14	100	60.31	22.03
WRF2.13.2	48	14	98	60.71	24.06
WRF2.14.2	43	14	110	72.70	20.05
WRF2.15.2	35	11	112	72.74	18.94
WRF2.16.2	43	12	111	71.79	19.52

Test-retest reliability. To examine test-retest reliability, we correlated student performance on the WRF forms that were administered during both the first and second sessions. Table 3 presents the results of these analyses. Overall, test-retest reliability was strong, ranging from .87 to .95.

Table 3
Test-retest Reliability of Grade 2 Word Reading Fluency Measures

Test Form	WRF2.11.2	WRF2.12.2	WRF2.13.2	WRF2.14.2	WRF2.15.2	WRF2.16.2
WRF2.11.1	0.94					
WRF2.12.1		0.95				
WRF2.13.1			0.93			
WRF2.14.1				0.89		
WRF2.15.1					0.92	
WRF2.16.1						0.87

Alternate form reliability. Alternate form reliability was evaluated using bivariate correlations among the different forms administered to students. Table 4 displays the results of these analyses. In general, we found strong positive relationships among the alternate forms, with correlations ranging from .92 to .95.

Table 4
Correlation between Alternate Forms of Grade 2 Word Reading Fluency Measures

Test Form	WRF2.12.1	WRF2.13.1	WRF2.15.1	WRF2.16.1
WRF2.11.1	0.95	0.92		
WRF2.12.1		0.95		
WRF2.14.1			0.94	0.92
WRF2.15.1				0.92

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 Word Reading Fluency analyses, 92, and 87% of the variance was associated with the 20, and 34 persons included in the analysis, 0% was associated with forms, and 0% was associated with occasion. The relative error variance was 6.37, and 11.16 for the first and second analysis, respectively. The absolute variance was 9.57 and 13.31, respectively. The G-Coefficients were .99 for the first analysis and .96 for the second analysis, while the phi coefficients were .98 and .96, respectively.

Word Reading Fluency: Forms 11, 12, & 13 (teachers 5 & 6)

Grade 2 WRF: Forms 11, 12 & 13

Generalizability ANOVA Table

Facet	*df*	SS	MS	Variance	Proportion
Persons	30	81439.26	2714.642	447.375	0.923
Forms	2	64.097	32.048	0.512	0.001
Occasions	1	581.941	581.941	6.048	0.012
Person*Forms	60	1144.903	19.082	0	0
Person*Occasion	30	1146.892	38.23	3.769	0.008
Forms*Occasion	2	16.333	8.167	0	0
Person*Forms*Occasions (Residual)	60	1615.333	26.922	26.922	0.056

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

Error Variances:

Relative, $\sigma^2(\delta_p)$	Absolute, $\sigma^2(\Delta_p)$
6.372	9.566

G-coefficients:

G: Ep^2	Phi: Φ
.986	.979

Grade 2 WRF: Forms 11, 12 & 13

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

n forms	*n* occasions				
	1	2	3	4	5
1	37.251	18.881	12.758	9.697	7.860
2	23.534	11.895	8.015	6.075	4.911
3	18.962	**9.566**	6.434	4.868	3.929
4	16.676	8.402	5.644	4.265	3.437
5	15.304	7.703	5.170	3.903	3.143

Grade 2 WRF: Forms 11, 12 & 13

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

n forms	*n* occasions				
	1	2	3	4	5
1	6.103	4.345	3.572	3.114	2.804
2	4.851	3.449	2.831	2.465	2.216
3	4.355	**3.093**	2.537	2.206	1.982
4	4.084	2.899	2.376	2.065	1.854
5	3.912	2.775	2.274	1.976	1.773

Grade 2 WRF: Forms 11, 12 & 13

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

n forms	*n* occasions				
	1	2	3	4	5
1	30.691	15.346	10.230	7.673	6.138
2	17.230	8.615	5.743	4.308	3.446
3	12.743	**6.372**	4.248	3.186	2.549
4	10.500	5.250	3.500	2.625	2.100
5	9.154	4.577	3.051	2.288	1.831

Grade 2 WRF: Forms 11, 12 & 13

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

n forms	*n* occasions				
	1	2	3	4	5
1	5.540	3.917	3.198	2.770	2.477
2	4.151	2.935	2.396	2.076	1.856
3	3.570	**2.524**	2.061	1.785	1.597
4	3.240	2.291	1.871	1.620	1.449
5	3.026	2.139	1.747	1.513	1.353

Grade 2 WRF: Forms 11, 12 & 13

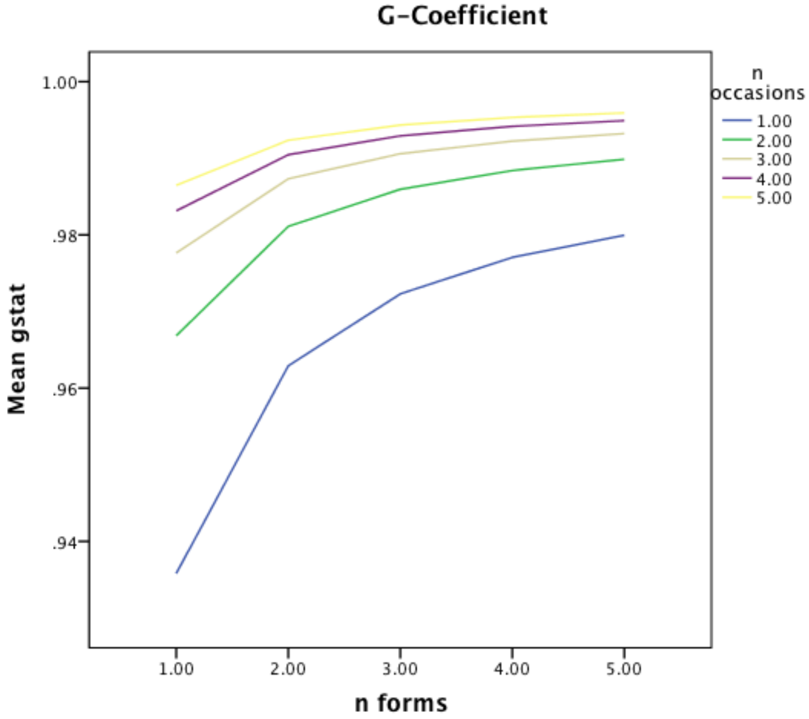
D-Study G Coefficients, Ep^2

n forms	*n* occasions				
	1	2	3	4	5
1	0.936	0.967	0.978	0.983	0.986
2	0.963	0.981	0.987	0.990	0.992
3	0.972	**0.986**	0.991	0.993	0.994
4	0.977	0.988	0.992	0.994	0.995
5	0.980	0.990	0.993	0.995	0.996

Grade 2 WRF: Forms 11, 12 & 13

D-Study Phi Coefficients, Φ

n forms	*n* occasions				
	1	2	3	4	5
1	0.923	0.960	0.972	0.979	0.983
2	0.950	0.974	0.982	0.987	0.989
3	0.959	**0.979**	0.986	0.989	0.991
4	0.964	0.982	0.988	0.991	0.992
5	0.967	0.983	0.989	0.991	0.993


G-Coefficient
Mean gstat
1.00
.98
.96
.94
n forms
1.00
2.00
3.00
4.00
5.00
n occasions
1.00
2.00
3.00
4.00
5.00

Word Reading Fluency: Forms 14, 15, 16 (teachers 7 & 8)

Grade 2 WRF: Forms 14, 15, & 16

Generalizability ANOVA Table

Facet	*df*	SS	MS	Variance	Proportion
Persons	33	62164.71	1883.779	302.807	0.868
Forms	2	260.51	130.255	0.911	0.003
Occasions	1	432.397	432.397	3.385	0.01
Person*Forms	66	2412.157	36.548	5.795	0.017
Person*Occasion	33	1826.436	55.347	10.129	0.029
Forms*Occasion	2	113.412	56.706	0.934	0.003
Person*Forms*Occasions (Residual)	66	1647.255	24.958	24.958	0.072

Note. Analysis included 34 students, with 3 forms (14, 15 & 16) on 2 occasions.

Error Variances:

Relative, $\sigma^2(\delta_p)$	Absolute, $\sigma^2(\Delta_p)$
11.156	13.308

G-coefficients:

G: Ep^2	Phi: Φ
.964	.958

Grade 2 WRF: Forms 14, 15, & 16

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

n forms	n occasions				
	1	2	3	4	5
1	46.113	26.409	19.842	16.558	14.587
2	29.814	16.583	12.173	9.968	8.645
3	24.381	**13.308**	9.617	7.772	6.664
4	21.664	11.670	8.339	6.673	5.674
5	20.034	10.688	7.572	6.014	5.080

Grade 2 WRF: Forms 14, 15, & 16

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

n forms	n occasions				
	1	2	3	4	5
1	6.791	5.139	4.454	4.069	3.819
2	5.460	4.072	3.489	3.157	2.940
3	4.938	**3.648**	3.101	2.788	2.581
4	4.654	3.416	2.888	2.583	2.382
5	4.476	3.269	2.752	2.452	2.254

Grade 2 WRF: Forms 14, 15, & 16

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

n forms	*n* occasions				
	1	2	3	4	5
1	40.883	23.339	17.491	14.567	12.812
2	25.506	14.202	10.434	8.550	7.419
3	20.380	**11.156**	8.081	6.544	5.621
4	17.818	9.633	6.905	5.541	4.722
5	16.280	8.719	6.199	4.939	4.183

Grade 2 WRF: Forms 14, 15, & 16

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

n forms	*n* occasions				
	1	2	3	4	5
1	6.394	4.831	4.182	3.817	3.579
2	5.050	3.769	3.230	2.924	2.724
3	4.514	**3.340**	2.843	2.558	2.371
4	4.221	3.104	2.628	2.354	2.173
5	4.035	2.953	2.490	2.222	2.045

Grade 2 WRF: Forms 14, 15, & 16

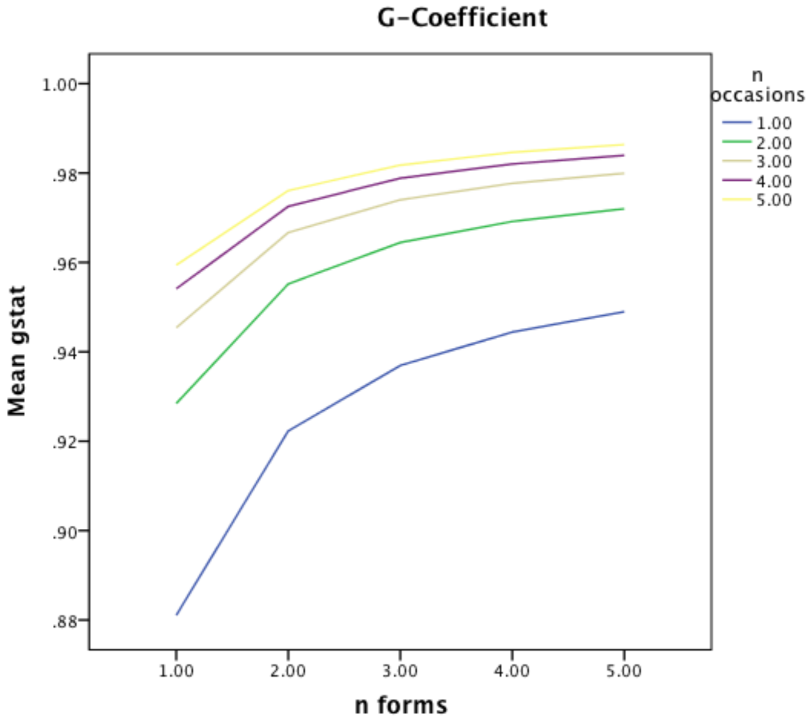
D-Study G Coefficients, Ep^2

n forms	*n* occasions				
	1	2	3	4	5
1	0.881	0.928	0.945	0.954	0.959
2	0.922	0.955	0.967	0.973	0.976
3	0.937	**0.964**	0.974	0.979	0.982
4	0.944	0.969	0.978	0.982	0.985
5	0.949	0.972	0.980	0.984	0.986

Grade 2 WRF: Forms 14, 15, & 16

D-Study Phi Coefficients, Φ

n forms	*n* occasions				
	1	2	3	4	5
1	0.868	0.920	0.939	0.948	0.954
2	0.910	0.948	0.961	0.968	0.972
3	0.925	**0.958**	0.969	0.975	0.978
4	0.933	0.963	0.973	0.978	0.982
5	0.938	0.966	0.976	0.981	0.984



Passage Reading Fluency

Descriptive statistics for the passage reading fluency measures are presented in Tables 5 and 6. Test-retest reliability results are presented in Table 7. Correlations between each of the 6 forms are presented in Table 8.

Table 5
Descriptive Statistics for Grade 2 Passage Reading Fluency Measures: Session 1

Test Form	*N*	Minimum	Maximum	Mean	Std. Deviation
PRF2.11.1	34	22	194	97.00	36.52
PRF2.12.1	34	25	151	87.62	33.49
PRF2.13.1	34	25	181	93.32	35.54
PRF2.14.1	50	18	223	120.32	41.06
PRF2.15.1	50	30	242	125.08	39.21
PRF2.16.1	50	17	203	114.08	37.76

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

Test Form	*N*	Minimum	Maximum	Mean	Std. Deviation
PRF2.11.2	48	19	194	97.79	41.05
PRF2.12.2	48	18	175	93.42	41.37
PRF2.13.2	48	24	181	97.52	40.20
PRF2.14.2	35	16	193	121.63	35.94
PRF2.15.2	43	23	225	129.88	37.86
PRF2.16.2	35	11	210	124.06	39.64

Test-retest reliability. To examine test-retest reliability, we correlated student performance on the PRF forms that were administered during both the first and second sessions. Table 7 presents the results of these analyses. Overall, test-retest reliability was strong, ranging from .88 to .96.

Table 7
Test-retest Reliability of Grade 2 Passage Reading Fluency Measures

Test Form	PRF2.11.2	PRF2.12.2	PRF2.13.2	PRF2.14.2	PRF2.15.2	PRF2.16.2
PRF2.11.1	0.88					
PRF2.12.1		0.96				
PRF2.13.1			0.93			
PRF2.14.1				0.90		
PRF2.15.1					0.94	
PRF2.16.1						0.95

Alternate form reliability. Alternate form reliability was evaluated using bivariate correlations among the different forms administered to students. Table 8 displays the results of these analyses. In general, we found strong positive relationships among the alternate forms, with correlations ranging from .91 to .95.

Table 8
Correlation between Alternate Forms of Grade 2 Passage Reading Fluency Measures

Test Form	PRF2.12.1	PRF2.13.1	PRF2.15.1	PRF2.16.1
PRF2.11.1	0.93	0.95		
PRF2.12.1		0.92		
PRF2.14.1			0.94	0.92
PRF2.15.1				0.91

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, 88% and 90% of the variance was associated with the 31 and 34 persons included in the analysis, 0% was associated with forms, and 0% was associated with occasion. The relative error variance was 39.03 and 25.54 for the first and second analysis, respectively. The absolute variance was 50.04 and 37.18, respectively. The G-Coefficients were .97 for the first analysis and .98 for the second analysis, while the phi coefficients were .96 and .97, respectively.

Passage Reading Fluency: Forms 11, 12, & 13 (teachers 5 & 6)

Grade 2 PRF: Forms 11, 12 & 13

Generalizability ANOVA Table

Facet	*df*	SS	MS	Variance	Proportion
Persons	30	242157.9	8071.932	1306.294	0.881
Forms	2	1329.849	664.925	5.872	0.004
Occasions	1	1865.167	1865.167	16.44	0.011
Person*Forms	60	8787.151	146.453	26.166	0.018
Person*Occasion	30	5455	181.833	29.237	0.02
Forms*Occasion	2	497.075	248.538	4.981	0.003
Person*Forms*Occasions (Residual)	60	5647.258	94.121	94.121	0.063

Note. Analysis included 31 students, with 3 forms (11, 12 & 13) on 2 occasions.

Error Variances:

Relative, $\sigma^2(\delta_p)$	Absolute, $\sigma^2(\Delta_p)$
39.027	50.035

G-coefficients:

G: Ep^2	Phi: Φ
.971	.963

Grade 2 PRF: Forms 11, 12 & 13

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

n forms	*n* occasions				
	1	2	3	4	5
1	176.817	104.427	80.297	68.233	60.994
2	111.247	63.633	47.762	39.826	35.065
3	89.391	**50.035**	36.916	30.357	26.422
4	78.462	43.236	31.494	25.623	22.100
5	71.905	39.156	28.240	22.782	19.507

Grade 2 PRF: Forms 11, 12 & 13

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

n forms	*n* occasions				
	1	2	3	4	5
1	13.297	10.219	8.961	8.260	7.810
2	10.547	7.977	6.911	6.311	5.922
3	9.455	**7.074**	6.076	5.510	5.140
4	8.858	6.575	5.612	5.062	4.701
5	8.480	6.257	5.314	4.773	4.417

Grade 2 PRF: Forms 11, 12 & 13

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

n forms	*n* occasions				
	1	2	3	4	5
1	149.524	87.845	67.285	57.005	50.837
2	89.381	51.232	38.516	32.157	28.342
3	69.333	**39.027**	28.926	23.875	20.844
4	59.309	32.925	24.131	19.733	17.095
5	53.295	29.264	21.254	17.249	14.845

Grade 2 PRF: Forms 11, 12 & 13

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

n forms	*n* occasions				
	1	2	3	4	5
1	12.228	9.373	8.203	7.550	7.130
2	9.454	7.158	6.206	5.671	5.324
3	8.327	**6.247**	5.378	4.886	4.566
4	7.701	5.738	4.912	4.442	4.135
5	7.300	5.410	4.610	4.153	3.853

Grade 2 PRF: Forms 11, 12 & 13

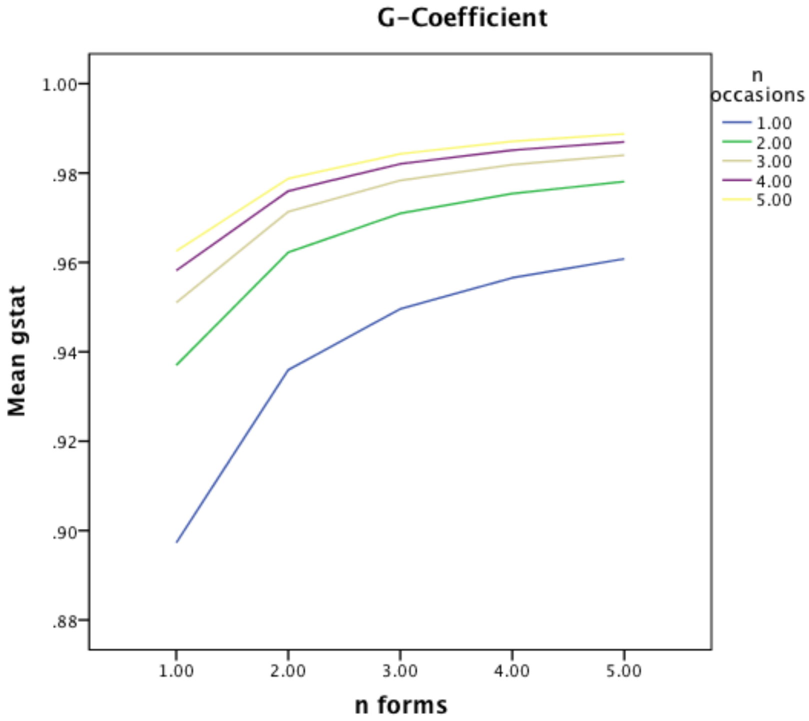
D-Study G Coefficients, Ep^2

n forms	*n* occasions				
	1	2	3	4	5
1	0.897	0.937	0.951	0.958	0.963
2	0.936	0.962	0.971	0.976	0.979
3	0.950	**0.971**	0.978	0.982	0.984
4	0.957	0.975	0.982	0.985	0.987
5	0.961	0.978	0.984	0.987	0.989

Grade 2 PRF: Forms 11, 12 & 13

D-Study Phi Coefficients, Φ

n forms	*n* occasions				
	1	2	3	4	5
1	0.881	0.926	0.942	0.950	0.955
2	0.922	0.954	0.965	0.970	0.974
3	0.936	**0.963**	0.973	0.977	0.980
4	0.943	0.968	0.976	0.981	0.983
5	0.948	0.971	0.979	0.983	0.985


G-Coefficient
Mean gstat
1.00
.98
.96
.94
.92
.90
.88
n forms
1.00
2.00
3.00
4.00
5.00
n occasions
1.00
2.00
3.00
4.00
5.00

Passage Reading Fluency: Forms 14, 15 & 16 (teachers 7 & 8)

Grade 2 PRF: Forms 14, 15 & 16

Generalizability ANOVA Table

Facet	*df*	SS	MS	Variance	Proportion
Persons	33	267579.8	8108.478	1325.873	0.901
Forms	2	3384.029	1692.015	15.603	0.011
Occasions	1	1445.338	1445.338	7.658	0.005
Person*Forms	66	6545.304	99.171	10.397	0.007
Person*Occasion	33	4370.828	132.449	18.024	0.012
Forms*Occasion	2	1220.382	610.191	15.642	0.011
Person*Forms*Occasions (Residual)	66	5172.951	78.378	78.378	0.053

Note. Analysis included 34 students, with 3 forms (14, 15 & 16) on 2 occasions.

Error Variances:

Relative, $\sigma^2(\delta_p)$	Absolute, $\sigma^2(\Delta_p)$
25.540	37.177

G-coefficients:

G: Ep^2	Phi: Φ
.981	.973

Grade 2 PRF: Forms 14, 15 & 16

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

n forms	*n* occasions				
	1	2	3	4	5
1	145.701	85.850	65.900	55.925	49.940
2	85.691	49.346	37.230	31.173	27.538
3	65.688	**37.177**	27.674	22.922	20.071
4	55.686	31.093	22.895	18.797	16.337
5	49.685	27.443	20.028	16.321	14.097

Grade 2 PRF: Forms 14, 15 & 16

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

n forms	*n* occasions				
	1	2	3	4	5
1	12.071	9.266	8.118	7.478	7.067
2	9.257	7.025	6.102	5.583	5.248
3	8.105	**6.097**	5.261	4.788	4.480
4	7.462	5.576	4.785	4.336	4.042
5	7.049	5.239	4.475	4.040	3.755

Grade 2 PRF: Forms 14, 15 & 16

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

n forms	*n* occasions				
	1	2	3	4	5
1	106.798	58.598	42.531	34.497	29.677
2	62.411	33.805	24.269	19.502	16.641
3	47.615	**25.540**	18.182	14.503	12.295
4	40.217	21.408	15.139	12.004	10.123
5	35.779	18.929	13.312	10.504	8.819

Grade 2 PRF: Forms 14, 15 & 16

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

n forms	*n* occasions				
	1	2	3	4	5
1	10.334	7.655	6.522	5.873	5.448
2	7.900	5.814	4.926	4.416	4.079
3	6.900	**5.054**	4.264	3.808	3.506
4	6.342	4.627	3.891	3.465	3.182
5	5.982	4.351	3.649	3.241	2.970

Grade 2 PRF: Forms 14, 15 & 16

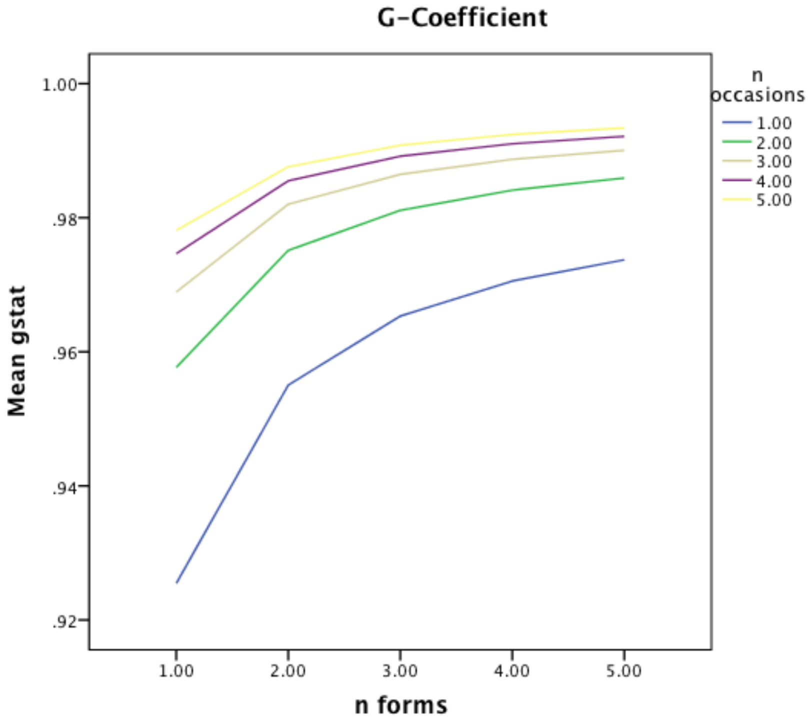
D-Study G Coefficients, Ep^2

n forms	*n* occasions				
	1	2	3	4	5
1	0.925	0.958	0.969	0.975	0.978
2	0.955	0.975	0.982	0.986	0.988
3	0.965	**0.981**	0.986	0.989	0.991
4	0.971	0.984	0.989	0.991	0.992
5	0.974	0.986	0.990	0.992	0.993

Grade 2 PRF: Forms 14, 15 & 16

D-Study Phi Coefficients, Φ

n forms	*n* occasions				
	1	2	3	4	5
1	0.901	0.939	0.953	0.960	0.964
2	0.939	0.964	0.973	0.977	0.980
3	0.953	**0.973**	0.980	0.983	0.985
4	0.960	0.977	0.983	0.986	0.988
5	0.964	0.980	0.985	0.988	0.989



Discussion

In this study, we examined the test-retest, alternate form reliability, and generalizability of two types of grade 2 easyCBM reading assessments. Both test-retest and alternate form reliability of word and passage reading fluency measures were positive and high. Correlations between the same form of these measures when administered one week apart and between alternate forms of these measures were found to be quite high. These findings add to the evidence of the technical adequacy of the grade 2 easyCBM reading measures.

The results of the G- and D-studies also increased the overall reliability evidence for the easyCBM reading measures. For the G-studies, the majority of variance was attributed to persons in every analysis, with 92% and 87% for Word Reading Fluency (WRF) and 88% and 90% for Passage Reading Fluency (PRF). The standard errors were also quite low. It is important to note that the error variances and dependability coefficients reported in text in the results section are those of the corresponding *analysis* and not of a particular form. For example, an examination of the error variance or standard error tables will show a bolded number, which is the error for the analysis. However, if only one form were given on one occasion then the error is increased (as reported in the D-study tables). Thus, in a classroom where decisions are made from one test form after one testing occasion, the error more closely resembles the one form on one occasion numbers reported in the D-study standard error tables.

Generally, increasing either facet (occasions or forms) resulted in a similar increase in the overall dependability. When examining the overall results, however, it is evident that using a single test form on a single occasion is generally sufficient for dependable measurement (i.e., > .8). This finding is important because other measurement systems have recommended using 3 fluency forms and taking the median score to increase reliability (Dibels*Next*, 2011) – a procedure that may appear unnecessary given the results of this study.

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.

Dibels*Next* (2011). *Dibels Oral Reading Fluency.* Retrieved February 14, 2011, from https://www.mclasshome.com/wgenhelp/dnext/DIBELS_Next/Assessment_and_Scoring/DORF_Details.htm

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

Test form administration order

Teacher	Word Reading Fluency		Passage Reading Fluency	
	Occasion 1	Occasion 2	Occasion 1	Occasion 2
5	13 – 12 – 11	13 – 11 – 12	13 – 12 – 11	13 – 11 – 12
6	11 – 12 – 13	12 – 13 – 11	11 – 12 – 13	12 – 13 – 11
7	16 – 15 – 14	16 – 14 – 15	16 – 15 – 14	16 – 14 – 15
8	14 – 15 – 16	15 – 16 – 14	14 – 15 – 16	15 – 16 – 14