Technical Report # 1216

An Examination of Test-Retest, Alternate Form Reliability, and Generalizability Theory Study of the easyCBM Reading Assessments: Grade 1

Daniel Anderson

Bitnara Jasmine Park

Cheng-Fei Lai

Julie Alonzo

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

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 1

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 researchers from the University of Oregon administered a battery of easyCBM assessments to students in participating classrooms. Data were gathered on two separate occasions, 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 control for order effects, with selected forms repeated across testing occasions 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

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

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 ($\mathrm{E}p^{2}$), 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. 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 two G-theory analyses for each of the passage reading fluency (PRF), word reading fluency (WRF), and letter sounds (LS) measure types, and three G-theory analyses for phoneme segmenting (PS). As Table 1 indicates, data from teacher 3 were missing for Occasion 1 across all measure types. Teacher 3 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 PRF analysis, data were collapsed for Teachers 1 and 2 and test forms 11 and 13 were examined. Form 12 was dropped from the analysis due to no administration occurring on occasion 1. The second analysis was identical but included only students instructed by teacher 4. Forms 14 and 16 were examined. For the first WRF analysis, data were dropped for teacher 1 because form 11 was not administered on occasion 2. Data for Teacher 2 were analyzed and test forms 11 and 12 were examined. The second WRF analysis included only teacher 4 with forms 11, 14, and 15 examined and forms 11 and 14 counterbalanced across occasions. For the first LS analysis, data were collapsed for Teacher 1 and 2 to examine the generalizability of forms 11 and 13. Form 12 was dropped from the analysis because it was not administered on occasion 1. Forms 14 and 16 were examined for teacher 4 for the second analysis. The PS design did not allow for any data to be collapsed across teachers. For the first analysis, forms 12 and 13 were examined for Teacher 1 in a non-counterbalanced design with form 11 dropped from the analysis. For the second analysis, forms 11 and 12 were examined for Teacher 2 in a counterbalanced design with form 13 dropped from the analysis. Forms 14, 15, and 16 were all examined for teacher 4 in a counterbalanced design for the third and final PS analysis.

For all g-theory analyses, forms were analyzed in ascending order regardless of administration order. For example, for the first analysis for PRF, the order of administration for forms 11 and 13 varied by the teacher and occasion. However, during the analysis the data were analyzed for forms 11 and 13 on the first occasion and forms 11 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 1 reading assessments are presented below, organized by type of measure.

Letter Sounds

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 1 Letter Sound Measures: Session 1

Test Form	*N*	Minimum	Maximum	Mean	Std. Deviation
LS1.11.1	42	29	65	44.45	9.37
LS1.13.1	42	28	90	47.24	13.17
LS1.14.1	20	22	86	55.15	16.47
LS1.16.1	20	31	74	49.50	13.50

Table 2
Descriptive Statistics for Grade 1 Letter Sound Measures: Session 2

Test Form	*N*	Minimum	Maximum	Mean	Std. Deviation
LS1.11.2	40	27	89	54.70	15.32
LS1.12.2	41	27	93	52.32	13.35
LS1.13.2	41	21	100	54.44	14.79
LS1.14.2	39	16	106	56.10	17.22
LS1.15.2	19	47	87	61.63	12.67
LS1.16.2	39	18	109	54.28	18.15

Test-retest reliability. To examine test-retest reliability, we correlated student performance on the LS forms that were administered during both the first and second sessions. Table 3 presents the results of these analyses. Overall, test-retest reliability was moderately strong, ranging from .77 to .87.

Table 3
Test-retest Reliability of Grade 1 Letter Sound Measures

Test Form	LS1.11.2	LS1.13.2	LS1.14.2	LS1.16.2
LS1.11.1	0.83			
LS1.13.1		0.86		
LS1.14.1			0.87	
LS1.16.1				0.77

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 moderately strong positive relationships among the alternate forms, with correlations ranging from .82 to .89.

Table 4
Correlation between Alternate Forms of Grade 1 Letter Sound Measures

Test Form	LS1.12.2	LS1.13.2	LS1.15.2	LS1.16.2
LS1.11.2	0.87	0.89		
LS1.12.2		0.83		
LS1.14.2			0.82	0.89
LS1.15.2				0.85

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 Letter Sounds analyses, 60% and 69% of the variance was associated with 15 and 37 *persons* included in the analysis, 0% and 5% were associated with *forms*, and 10% and 16% were associated with *occasion*. The relative error variance was 10.38 for the first analysis and 20.31 for the second while the absolute variance 30.22 and 39.14 respectively. The G-Coefficients were .87 for the first analysis and .95 for the second, while the phi coefficients were .87 and .95, respectively.

Letter Sounds: Forms 11 & 13 (Teachers 1 & 2)

Grade 1 LS: Forms 11 & 13

Generalizability ANOVA Table

Facet	*df*	SS	MS	Variance	Proportion
Persons	36	22525.203	625.7	136.114	0.603
Forms	1	60.98	60.98	0.000	0.000
Occasions	1	2845.953	2845.953	36.369	0.161
Person*Forms	36	1532.77	42.577	11.18	0.05
Person*Occasion	36	2119.797	58.883	19.333	0.086
Forms*Occasion	1	115.953	115.953	2.587	0.011
Person*Forms*Occasions (Residual)	36	727.797	20.217	20.217	0.09

Note. Analysis included 37 students, with 2 forms (11 & 13) on 2 occasions.

Error Variances:

Relative, $\sigma^2(\delta_p)$	Absolute, $\sigma^2(\Delta_p)$
20.311	39.142

G-coefficients:

G: Ep^2	Phi: Φ
.870	.777

Grade 1 LS: Forms 11 & 13

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

n forms	*n* occasions				
	1	2	3	4	5
1	89.687	50.434	37.349	30.807	26.882
2	72.695	**39.142**	27.958	22.366	19.011
3	67.031	35.379	24.828	19.553	16.388
4	64.199	33.497	23.263	18.146	15.076
5	62.5	32.368	22.324	17.302	14.289

Grade 1 LS: Forms 11 & 13

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

n forms	*n* occasions				
	1	2	3	4	5
1	9.470	7.102	6.111	5.550	5.185
2	8.526	**6.256**	5.288	4.729	4.360
3	8.187	5.948	4.983	4.422	4.048
4	8.012	5.788	4.823	4.260	3.883
5	7.906	5.689	4.725	4.160	3.780

Grade 1 LS: Forms 11 & 13

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

n forms	*n* occasions				
	1	2	3	4	5
1	50.73	30.955	24.363	21.068	19.09
2	35.032	**20.311**	15.404	12.95	11.478
3	29.799	16.763	12.417	10.245	8.941
4	27.183	14.989	10.924	8.892	7.673
5	25.613	13.924	10.028	8.08	6.911

Grade 1 LS: Forms 11 & 13

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

n forms	*n* occasions				
	1	2	3	4	5
1	7.122	5.564	4.936	4.590	4.369
2	5.919	**4.507**	3.925	3.599	3.388
3	5.459	4.094	3.524	3.201	2.990
4	5.214	3.872	3.305	2.982	2.770
5	5.061	3.731	3.167	2.843	2.629

Grade 1 LS: Forms 11 & 13

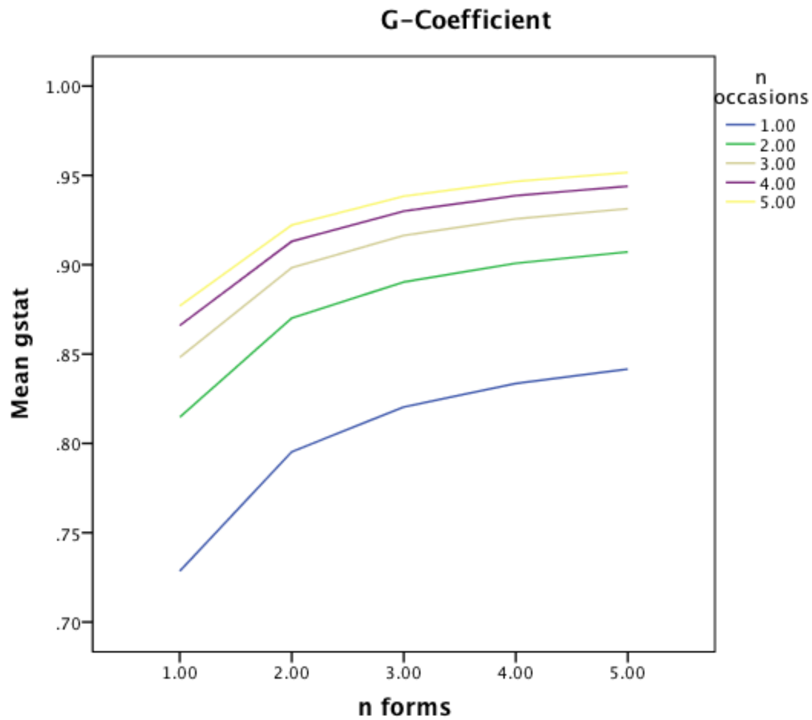
D-Study G Coefficients, Ep^2

n forms	*n* occasions				
	1	2	3	4	5
1	0.728	0.815	0.848	0.866	0.877
2	0.795	**0.870**	0.898	0.913	0.922
3	0.820	0.890	0.916	0.93	0.938
4	0.834	0.901	0.926	0.939	0.947
5	0.842	0.907	0.931	0.944	0.952

Grade 1 LS: Forms 11 & 13

D-Study Phi Coefficients, Φ

n forms	*n* occasions				
	1	2	3	4	5
1	0.603	0.730	0.785	0.815	0.835
2	0.652	**0.777**	0.830	0.859	0.877
3	0.670	0.794	0.846	0.874	0.893
4	0.680	0.803	0.854	0.882	0.900
5	0.685	0.808	0.859	0.887	0.905


G-Coefficient
Mean gstat
1.00
.95
.90
.85
.80
.75
.70
1.00
2.00
3.00
4.00
5.00
n forms
n occasions
1.00
2.00
3.00
4.00
5.00

Letter Sounds: Forms 14 & 16 (Teacher 4)

Grade 1 LS: Forms 14 & 16

Generalizability ANOVA Table

Facet	*df*	SS	MS	Variance	Proportion
Persons	14	10679.733	762.838	181.514	0.694
Forms	1	385.067	385.067	12.938	0.049
Occasions	1	806.667	806.667	26.748	0.102
Person*Forms	14	478.933	34.21	0.000	0.000
Person*Occasion	14	581.333	41.524	1.286	0.005
Forms*Occasion	1	1.667	1.667	0.000	0.000
Person*Forms*Occasions (Residual)	14	545.333	38.952	38.952	0.149

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

Error Variances:

Relative, $\sigma^2(\delta_p)$	Absolute, $\sigma^2(\Delta_p)$
10.381	30.224

G-coefficients:

G: Ep^2	Phi: Φ
.946	.857

Grade 1 LS: Forms 14 & 16

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

n forms	*n* occasions				
	1	2	3	4	5
1	79.924	46.431	35.267	29.685	26.335
2	53.979	**30.224**	22.306	18.346	15.971
3	45.33	24.821	17.985	14.567	12.516
4	41.006	22.120	15.825	12.677	10.789
5	38.411	20.500	14.529	11.544	9.752

Grade 1 LS: Forms 14 & 16

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

n forms	*n* occasions				
	1	2	3	4	5
1	8.940	6.814	5.939	5.448	5.132
2	7.347	**5.498**	4.723	4.283	3.996
3	6.733	4.982	4.241	3.817	3.538
4	6.404	4.703	3.978	3.560	3.285
5	6.198	4.528	3.812	3.398	3.123

Grade 1 LS: Forms 14 & 16

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

n forms	*n* occasions				
	1	2	3	4	5
1	40.238	20.119	13.413	10.06	8.048
2	20.762	**10.381**	6.921	5.19	4.152
3	14.27	7.135	4.757	3.567	2.854
4	11.024	5.512	3.675	2.756	2.205
5	9.076	4.538	3.025	2.269	1.815

Grade 1 LS: Forms 14 & 16

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

n forms	*n* occasions				
	1	2	3	4	5
1	6.343	4.485	3.662	3.172	2.837
2	4.557	**3.222**	2.631	2.278	2.038
3	3.778	2.671	2.181	1.889	1.689
4	3.320	2.348	1.917	1.660	1.485
5	3.013	2.130	1.739	1.506	1.347

Grade 1 LS: Forms 14 & 16

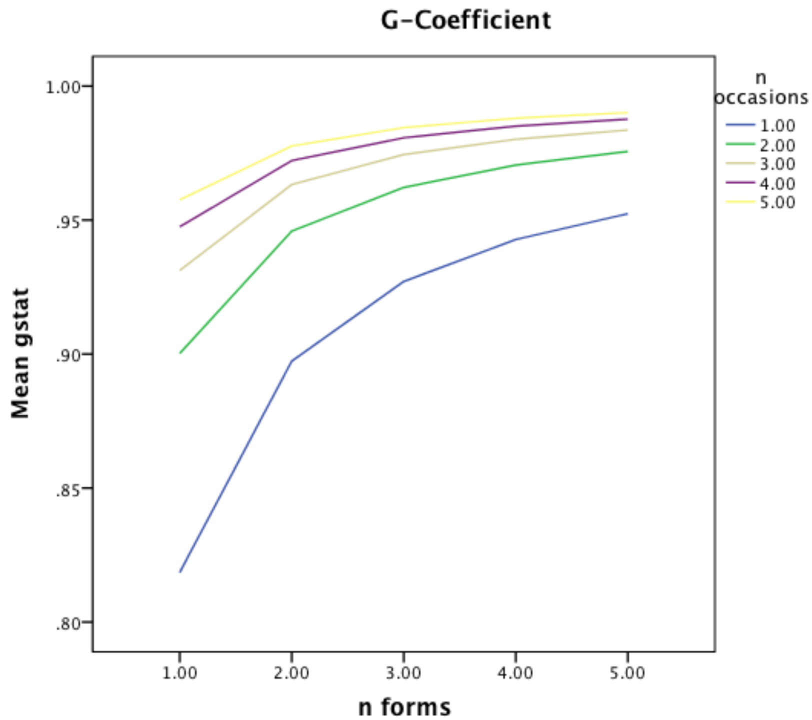
D-Study G Coefficients, Ep^2

n forms	*n* occasions				
	1	2	3	4	5
1	0.819	0.900	0.931	0.947	0.958
2	0.897	**0.946**	0.963	0.972	0.978
3	0.927	0.962	0.974	0.981	0.985
4	0.943	0.971	0.98	0.985	0.988
5	0.952	0.976	0.984	0.988	0.99

Grade 1 LS: Forms 14 & 16

D-Study Phi Coefficients, Φ

n forms	*n* occasions				
	1	2	3	4	5
1	0.694	0.796	0.837	0.859	0.873
2	0.771	**0.857**	0.891	0.908	0.919
3	0.8	0.880	0.91	0.926	0.935
4	0.816	0.891	0.92	0.935	0.944
5	0.825	0.899	0.926	0.94	0.949



Phoneme Segmenting

Descriptive statistics 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 1 Segmenting Measures: Session 1

Test Form	*N*	Minimum	Maximum	Mean	Std. Deviation
Seg1.11.1	42	20	65	48.81	9.67
Seg1.12.1	42	20	64	50.12	9.62
Seg1.13.1	42	19	68	49.71	9.19
Seg1.14.1	22	37	69	55.09	9.79
Seg1.15.1	21	34	70	57.05	9.43
Seg1.16.1	22	30	70	53.82	9.37

Table 6
Descriptive Statistics for Grade 1 Phoneme Segmenting Measures: Session 2

Test Form	*N*	Minimum	Maximum	Mean	Std. Deviation
Seg1.11.2	20	39	88	57.25	13.00
Seg1.12.2	40	21	88	54.28	12.10
Seg1.13.2	20	24	64	49.75	9.94
Seg1.14.2	39	28	73	51.23	11.53
Seg1.15.2	39	27	73	51.49	11.40
Seg1.16.2	19	29	58	46.11	8.69

Test-retest reliability. To examine test-retest reliability, we correlated student performance on the phoneme segmenting forms that were administered during both the first and second sessions. Table 7 presents the results of these analyses. Overall, test-retest reliability was moderate, ranging from .50 to .81. The test-retest reliability of Form 16 was not statistically significant, most likely due to the small sample size (n=15).

Table 7
Test-retest Reliability of Grade 1 Segmenting Measures

Test Form	Seg1.11.2	Seg1.12.2	Seg1.13.2	Seg1.14.2	Seg1.15.2	Seg1.16.2
Seg1.11.1	0.56					
Seg1.12.1		0.50				
Seg1.13.1			0.81			
Seg1.14.1				0.58		
Seg1.15.1					0.72	
Seg1.16.1						0.32*

* $p > .05$.

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 moderately strong positive relationships among the alternate forms, with correlations ranging from .62 to .89.

Table 8
Correlation between Alternate Forms of Grade 1 Phoneme Segmenting Measures

Test Form	Seg1.12.2	Seg1.13.2	Seg1.15.2	Seg1.16.2
Seg1.11.2	0.89			
Seg1.12.2		0.82		
Seg1.13.2				
Seg1.14.2			0.78	0.62
Seg1.15.2				0.67

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 phoneme segmenting analyses, 29-60% of the variance was associated with the 15-18 persons included in the analysis, 0-2% was associated with forms, and 1-9% was associated with occasion. There was a quite large interaction between persons and order, ranging from 11-48% of the total variance. The relative error variance ranged from 15.89-32.44, while the absolute variance ranged from 16.64 to 35.38. The G-Coefficients ranged from .50-.83, while the phi coefficients ranged from .47-.82.

Phoneme Segmenting: Forms 12 & 13 (Teacher 1)

Grade 1 PS: Forms 12 & 13

Generalizability ANOVA Table

Facet	*df*	SS	MS	Variance	Proportion
Persons	17	6290.500	370.029	77.714	0.599
Forms	1	8.000	8.000	0.000	0.000
Occasions	1	117.556	117.556	0.819	0.006
Person*Forms	17	537.000	31.588	0.000	0.000
Person*Occasion	17	1080.444	63.556	13.792	0.106
Forms*Occasion	1	60.500	60.500	1.363	0.011
Person*Forms*Occasions (Residual)	17	611.500	35.971	35.971	0.277

Note. Analysis included 18 students, with 2 forms (12 & 13) on 2 occasions.

Error Variances:

Relative, $\sigma^2(\delta_p)$	Absolute, $\sigma^2(\Delta_p)$
15.889	16.639

G-coefficients:

G: Eρ^2	Phi: Φ
.830	.824

Grade 1 PS: Forms 12 & 13

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

n forms	*n* occasions				
	1	2	3	4	5
1	51.944	25.972	17.315	12.986	10.389
2	33.278	**16.639**	11.093	8.319	6.656
3	27.056	13.528	9.019	6.764	5.411
4	23.944	11.972	7.981	5.986	4.789
5	22.078	11.039	7.359	5.519	4.416

Grade 1 PS: Forms 12 & 13

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

n forms	*n* occasions				
	1	2	3	4	5
1	7.207	5.096	4.161	3.604	3.223
2	5.769	**4.079**	3.331	2.884	2.580
3	5.202	3.678	3.003	2.601	2.326
4	4.893	3.460	2.825	2.447	2.188
5	4.699	3.322	2.713	2.349	2.101

Grade 1 PS: Forms 12 & 13

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

n forms	*n* occasions				
	1	2	3	4	5
1	49.763	24.882	16.588	12.441	9.953
2	31.778	**15.889**	10.593	7.944	6.356
3	25.783	12.891	8.594	6.446	5.157
4	22.785	11.393	7.595	5.696	4.557
5	20.987	10.493	6.996	5.247	4.197

Grade 1 PS: Forms 12 & 13

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

n forms	*n* occasions				
	1	2	3	4	5
1	7.054	4.988	4.073	3.527	3.155
2	5.637	**3.986**	3.255	2.819	2.521
3	5.078	3.590	2.932	2.539	2.271
4	4.773	3.375	2.756	2.387	2.135
5	4.581	3.239	2.645	2.291	2.049

Grade 1 PS: Forms 12 & 13

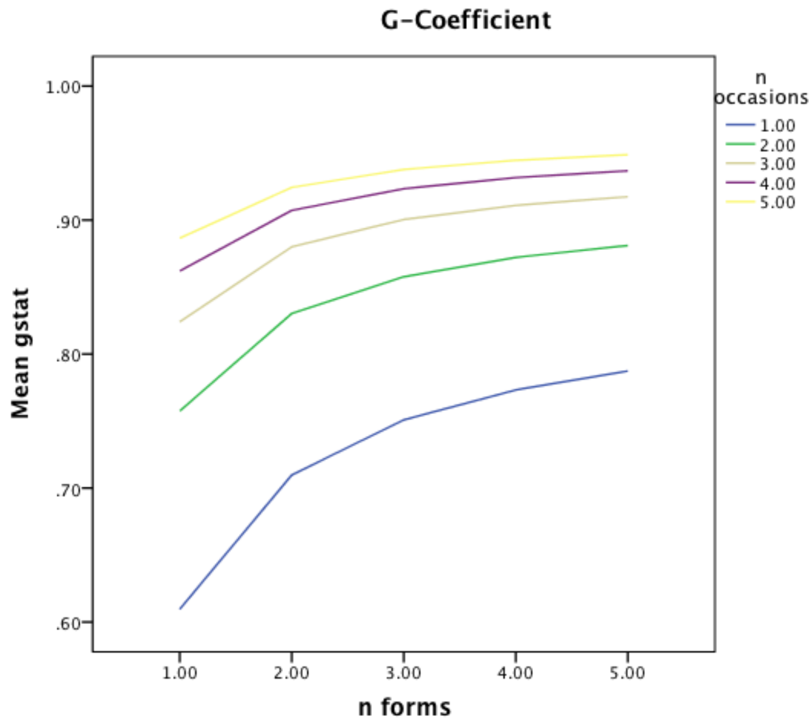
D-Study G Coefficients, Ep^2

n forms	*n* occasions				
	1	2	3	4	5
1	0.610	0.757	0.824	0.862	0.886
2	0.710	**0.830**	0.880	0.907	0.924
3	0.751	0.858	0.900	0.923	0.938
4	0.773	0.872	0.911	0.932	0.945
5	0.787	0.881	0.917	0.937	0.949

Grade 1 PS: Forms 12 & 13

D-Study Phi Coefficients, Φ

n forms	*n* occasions				
	1	2	3	4	5
1	0.599	0.750	0.818	0.857	0.882
2	0.700	**0.824**	0.875	0.903	0.921
3	0.742	0.852	0.896	0.920	0.935
4	0.764	0.867	0.907	0.928	0.942
5	0.779	0.876	0.913	0.934	0.946



Phoneme Segmenting: Forms 11 & 12 (Teacher 2)

Grade 1 PS: Forms 11 & 12

Generalizability ANOVA Table

Facet	*df*	SS	MS	Variance	Proportion
Persons	17	4370.569	257.092	31.83	0.293
Forms	1	45.125	45.125	0.873	0.008
Occasions	1	284.014	284.014	5.005	0.046
Person*Forms	17	441.125	25.949	6.85	0.063
Person*Occasion	17	1973.236	116.073	51.912	0.477
Forms*Occasion	1	0.014	0.014	0.000	0.000
Person*Forms*Occasions (Residual)	17	208.236	12.249	12.249	0.113

Note. Analysis included 18 students, with 2 forms (11 & 12) on 2 occasions.

Error Variances:

Relative, $\sigma^2(\delta_p)$	Absolute, $\sigma^2(\Delta_p)$
32.443	35.382

G-coefficients:

G: Ep^2	Phi: Φ
.495	.474

Grade 1 PS: Forms 11 & 12

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

n forms	*n* occasions				
	1	2	3	4	5
1	76.888	42.305	30.778	25.014	21.555
2	66.902	**35.382**	24.875	19.621	16.469
3	63.574	33.074	22.907	17.824	14.774
4	61.91	31.92	21.924	16.925	13.926
5	60.911	31.228	21.333	16.386	13.418

Grade 1 PS: Forms 11 & 12

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

n forms	*n* occasions				
	1	2	3	4	5
1	8.769	6.504	5.548	5.001	4.643
2	8.179	**5.948**	4.987	4.430	4.058
3	7.973	5.751	4.786	4.222	3.844
4	7.868	5.650	4.682	4.114	3.732
5	7.805	5.588	4.619	4.048	3.663

Grade 1 PS: Forms 11 & 12

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

n forms	*n* occasions				
	1	2	3	4	5
1	71.011	38.93	28.237	22.89	19.682
2	61.461	**32.443**	22.77	17.934	15.032
3	58.278	30.281	20.948	16.282	13.482
4	56.686	29.199	20.037	15.456	12.707
5	55.732	28.551	19.49	14.96	12.242

Grade 1 PS: Forms 11 & 12

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

n forms	*n* occasions				
	1	2	3	4	5
1	8.427	6.239	5.314	4.784	4.436
2	7.840	**5.696**	4.772	4.235	3.877
3	7.634	5.503	4.577	4.035	3.672
4	7.529	5.404	4.476	3.931	3.565
5	7.465	5.343	4.415	3.868	3.499

Grade 1 PS: Forms 11 & 12

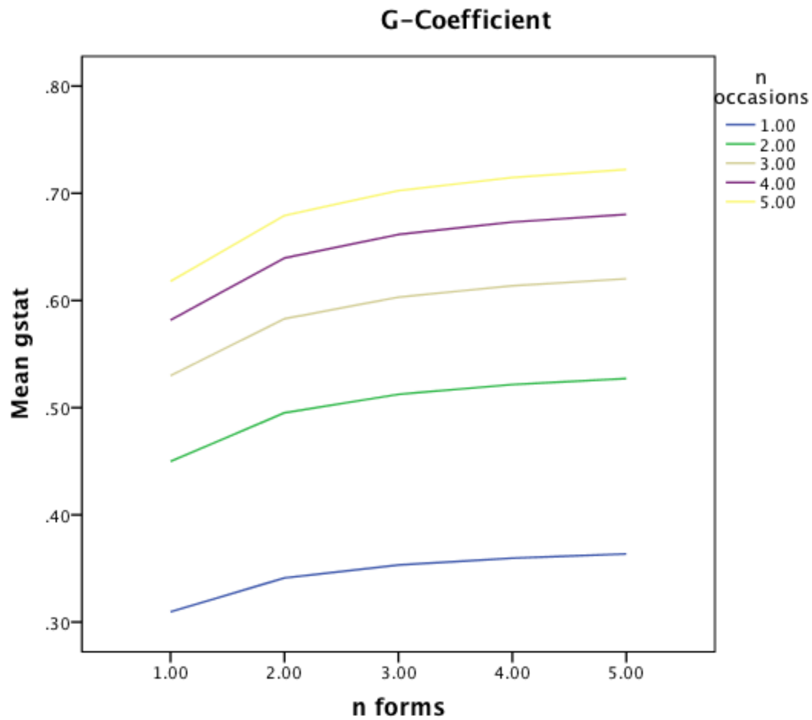
D-Study G Coefficients, Ep^2

n forms	*n* occasions				
	1	2	3	4	5
1	0.31	0.45	0.53	0.582	0.618
2	0.341	**0.495**	0.583	0.64	0.679
3	0.353	0.512	0.603	0.662	0.702
4	0.36	0.522	0.614	0.673	0.715
5	0.364	0.527	0.62	0.68	0.722

Grade 1 PS: Forms 11 & 12

D-Study Phi Coefficients, Φ

n forms	*n* occasions				
	1	2	3	4	5
1	0.293	0.429	0.508	0.56	0.596
2	0.322	**0.474**	0.561	0.619	0.659
3	0.334	0.49	0.582	0.641	0.683
4	0.34	0.499	0.592	0.653	0.696
5	0.343	0.505	0.599	0.66	0.703



Phoneme Segmenting: Forms 14, 15, & 16 (Teacher 4)

Grade 1 PS: Forms 14, 15, & 16

Generalizability ANOVA Table

Facet	*df*	SS	MS	Variance	Proportion
Persons	14	4900.489	350.035	41.979	0.407
Forms	2	226.956	113.478	2.021	0.020
Occasions	1	547.6	547.6	9.587	0.093
Person*Forms	28	783.711	27.99	3.401	0.033
Person*Occasion	14	1279.067	91.362	23.391	0.227
Forms*Occasion	2	92.067	46.033	1.656	0.016
Person*Forms*Occasions (Residual)	28	593.267	21.188	21.188	0.205

Note. Analysis included 15 students, with 3 forms (11 & 13) on 2 occasions.

Error Variances:

Relative, $\sigma^2(\delta_p)$	Absolute, $\sigma^2(\Delta_p)$
16.361	22.104

G-coefficients:

G: Ep^2	Phi: Φ
.720	.655

Grade 1 PS: Forms 14, 15, & 16

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

n forms	*n* occasions				
	1	2	3	4	5
1	61.244	33.333	24.03	19.378	16.587
2	47.111	24.911	17.511	13.811	11.591
3	42.4	**22.104**	15.338	11.956	9.926
4	40.044	20.7	14.252	11.028	9.093
5	38.631	19.858	13.6	10.471	8.594

Grade 1 PS: Forms 14, 15, & 16

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

n forms	*n* occasions				
	1	2	3	4	5
1	7.826	5.773	4.902	4.402	4.073
2	6.864	4.991	4.185	3.716	3.405
3	6.512	**4.701**	3.916	3.458	3.151
4	6.328	4.550	3.775	3.321	3.015
5	6.215	4.456	3.688	3.236	2.932

Grade 1 PS: Forms 14, 15, & 16

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

n forms	*n* occasions				
	1	2	3	4	5
1	47.98	25.69	18.261	14.546	12.317
2	35.686	18.693	13.029	10.197	8.497
3	31.588	**16.361**	11.285	8.747	7.224
4	29.538	15.194	10.413	8.022	6.588
5	28.309	14.495	9.89	7.587	6.206

Grade 1 PS: Forms 14, 15, & 16

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

n forms	*n* occasions				
	1	2	3	4	5
1	6.927	5.069	4.273	3.814	3.510
2	5.974	4.324	3.610	3.193	2.915
3	5.620	**4.045**	3.359	2.958	2.688
4	5.435	3.898	3.227	2.832	2.567
5	5.321	3.807	3.145	2.754	2.491

Grade 1 PS: Forms 14, 15, & 16

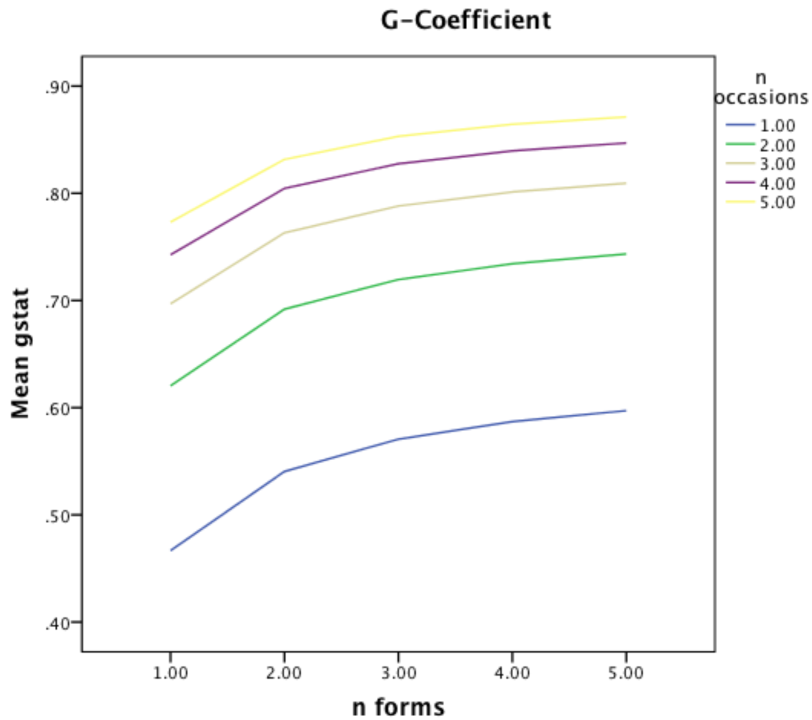
D-Study G Coefficients, Ep^2

n forms	*n* occasions				
	1	2	3	4	5
1	0.467	0.62	0.697	0.743	0.773
2	0.541	0.692	0.763	0.805	0.832
3	0.571	**0.72**	0.788	0.828	0.853
4	0.587	0.734	0.801	0.84	0.864
5	0.597	0.743	0.809	0.847	0.871

Grade 1 PS: Forms 14, 15, & 16

D-Study Phi Coefficients, Φ

n forms	*n* occasions				
	1	2	3	4	5
1	0.407	0.557	0.636	0.684	0.717
2	0.471	0.628	0.706	0.752	0.784
3	0.498	**0.655**	0.732	0.778	0.809
4	0.512	0.67	0.747	0.792	0.822
5	0.521	0.679	0.755	0.8	0.83



Word Reading Fluency

Descriptive statistics are presented in Tables 9 and 10. Test-retest reliability results are presented in Table 11. Correlations between each of the 4 forms are presented in Table 12.

Table 9
Descriptive Statistics for Grade 1 Word Reading Fluency Measures: Session 1

Test Form	*N*	Minimum	Maximum	Mean	Std. Deviation
WRF1.11.1	62	11	114	54.89	25.94
WRF1.12.1	42	11	140	49.76	26.59
WRF1.14.1	20	39	104	64.50	16.34
WRF1.15.1	20	35	98	64.50	17.21

Table 10
Descriptive Statistics for Grade 1 Word Reading Fluency Measures: Session 2

Test Form	*N*	Minimum	Maximum	Mean	Std. Deviation
WRF1.11.2	41	13	124	61.24	27.50
WRF1.12.2	60	11	110	50.48	24.50
WRF1.14.2	20	41	98	68.25	16.96
WRF1.15.2	20	44	96	65.95	15.02

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 11 presents the results of these analyses. Overall, test-retest reliability was strong, ranging from .87 to .95.

Table 11
Test-retest Reliability of Grade 1 Word Reading Fluency Measures

Test Form	WRF1.11.2	WRF1.12.2	WRF1.14.2	WRF1.15.2
WRF1.11.1	0.93			
WRF1.12.1		0.95		
WRF1.14.1			0.87	
WRF1.15.1				0.91

Alternate form reliability. Alternate form reliability was evaluated using bivariate correlations among the different forms administered to students. Table 12 displays the results of these analyses. In general, we found strong positive relationships among the alternate forms, with correlations ranging from .89 to .97.

Table 12
Correlation between Alternate Forms of Grade 1 Word Reading Fluency Measures

Test Form	WRF1.12.1	WRF1.14.1	WRF1.15.1
WRF1.11.1	0.97	0.91	0.95
WRF1.14.1			0.89

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 Word Reading Fluency analyses, 94% and 85% of the variance was associated with the 19 and 15 persons included in the analysis, 0% was associated with forms, and 0% was associated with occasion. The relative error variance was 15.13 for the first analysis and 13.00 for the second, while the absolute variance was 18.03 and 14.61 respectively. The G-Coefficients were .98 for the first analysis and .96 for the second, while the phi coefficients were .98 and .95 respectively.

Word Reading Fluency: Forms 11 & 12 (teacher 2)

Grade 1 WRF: Forms 11 & 12

Generalizability ANOVA Table

Facet	*df*	SS	MS	Variance	Proportion
Persons	18	61565.79	3420.322	844.99	0.935
Forms	1	32.895	32.895	0	0
Occasions	1	280.474	280.474	1.232	0.001
Person*Forms	18	281.105	15.617	0	0
Person*Occasion	18	1089.526	60.529	12.373	0.014
Forms*Occasion	1	208.895	208.895	9.111	0.01
Person*Forms*Occasions (Residual)	18	644.105	35.784	35.784	0.04

Note. Analysis included 19 students, with 2 forms (11 & 12) on 2 occasions.

Error Variances:

Relative, $\sigma^2(\delta_p)$	Absolute, $\sigma^2(\Delta_p)$
15.132	18.026

G-coefficients:

G: Eρ^2	Phi: Φ
.982	.979

Grade 1 WRF: Forms 11 & 12

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

n forms	*n* occasions				
	1	2	3	4	5
1	58.5	29.25	19.5	14.625	11.7
2	36.053	**18.026**	12.018	9.013	7.211
3	28.57	14.285	9.523	7.143	5.714
4	24.829	12.414	8.276	6.207	4.966
5	22.584	11.292	7.528	5.646	4.517

Grade 1 WRF: Forms 11 & 12

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

n forms	*n* occasions				
	1	2	3	4	5
1	7.649	5.408	4.416	3.824	3.421
2	6.004	**4.246**	3.467	3.002	2.685
3	5.345	3.780	3.086	2.673	2.390
4	4.983	3.523	2.877	2.491	2.228
5	4.752	3.360	2.744	2.376	2.125

Grade 1 WRF: Forms 11 & 12

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

n forms	*n* occasions				
	1	2	3	4	5
1	48.156	24.078	16.052	12.039	9.631
2	30.265	**15.132**	10.088	7.566	6.053
3	24.301	12.15	8.1	6.075	4.86
4	21.319	10.659	7.106	5.33	4.264
5	19.53	9.765	6.51	4.882	3.906

Grade 1 WRF: Forms 11 & 12

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

n forms	*n* occasions				
	1	2	3	4	5
1	6.939	4.907	4.006	3.470	3.103
2	5.501	**3.890**	3.176	2.751	2.460
3	4.930	3.486	2.846	2.465	2.205
4	4.617	3.265	2.666	2.309	2.065
5	4.419	3.125	2.551	2.210	1.976

Grade 1 WRF: Forms 11 & 12

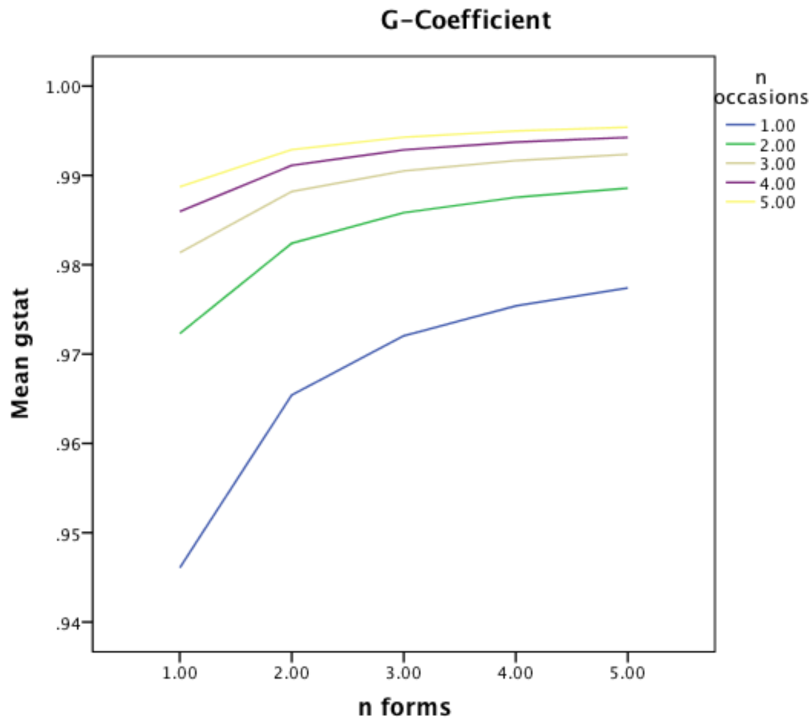
D-Study G Coefficients, Ep^2

n forms	*n* occasions				
	1	2	3	4	5
1	0.946	0.972	0.981	0.986	0.989
2	0.965	**0.982**	0.988	0.991	0.993
3	0.972	0.986	0.991	0.993	0.994
4	0.975	0.988	0.992	0.994	0.995
5	0.977	0.989	0.992	0.994	0.995

Grade 1 WRF: Forms 11 & 12

D-Study Phi Coefficients, Φ

n forms	*n* occasions				
	1	2	3	4	5
1	0.935	0.967	0.977	0.983	0.986
2	0.959	**0.979**	0.986	0.989	0.992
3	0.967	0.983	0.989	0.992	0.993
4	0.971	0.986	0.99	0.993	0.994
5	0.974	0.987	0.991	0.993	0.995



G-Coefficient
Mean gstat
1.00
.99
.98
.97
.96
.95
.94
1.00
2.00
3.00
4.00
5.00
n forms
n occasions
1.00
2.00
3.00
4.00
5.00

Word Reading Fluency: Forms 11, 14, and 15 (teacher 4)

Grade 1 WRF: Forms 11, 14 & 15
Generalizability ANOVA Table

Facet	*df*	SS	MS	Variance	Proportion
Persons	14	25123.29	1794.521	286.089	0.847
Forms	2	374.689	187.344	0.816	0.002
Occasions	1	11.378	11.378	0.000	0.000
Person*Forms	28	1172.311	41.868	11.162	0.033
Person*Occasion	14	779.289	55.663	12.04	0.036
Forms*Occasion	2	281.089	140.544	8.067	0.024
Person*Forms*Occasions (Residual)	28	547.244	19.544	19.544	0.058

Note. Analysis included 15 students, with 3 forms (11, 14 & 15) on 2 occasions.

Error Variances:

Relative, $\sigma^2(\delta_p)$	Absolute, $\sigma^2(\Delta_p)$
12.998	14.614

G-coefficients:

G: Ep^2	Phi: Φ
.957	.951

Grade 1 WRF: Forms 11, 14 & 15

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

n forms	*n* occasions				
	1	2	3	4	5
1	51.629	31.803	25.195	21.89	19.908
2	31.834	18.912	14.604	12.45	11.158
3	25.236	**14.614**	11.074	9.303	8.241
4	21.937	12.466	9.309	7.73	6.783
5	19.957	11.177	8.25	6.786	5.908

Grade 1 WRF: Forms 11, 14 & 15

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

n forms	*n* occasions				
	1	2	3	4	5
1	7.185	5.639	5.019	4.679	4.462
2	5.642	**4.349**	3.822	3.528	3.340
3	5.024	3.823	3.328	3.050	2.871
4	4.684	3.531	3.051	2.780	2.604
5	4.467	3.343	2.872	2.605	2.431

Grade 1 WRF: Forms 11, 14 & 15

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

n forms	*n* occasions				
	1	2	3	4	5
1	42.746	26.954	21.69	19.058	17.479
2	27.393	16.487	12.852	11.034	9.943
3	22.275	**12.998**	9.905	8.359	7.432
4	19.716	11.253	8.432	7.022	6.176
5	18.181	10.207	7.549	6.22	5.422

Grade 1 WRF: Forms 11, 14 & 15

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

n forms	*n* occasions				
	1	2	3	4	5
1	6.538	5.192	4.657	4.366	4.181
2	5.234	4.060	3.585	3.322	3.153
3	4.720	**3.605**	3.147	2.891	2.726
4	4.440	3.355	2.904	2.650	2.485
5	4.264	3.195	2.748	2.494	2.329

Grade 1 WRF: Forms 11, 14 & 15

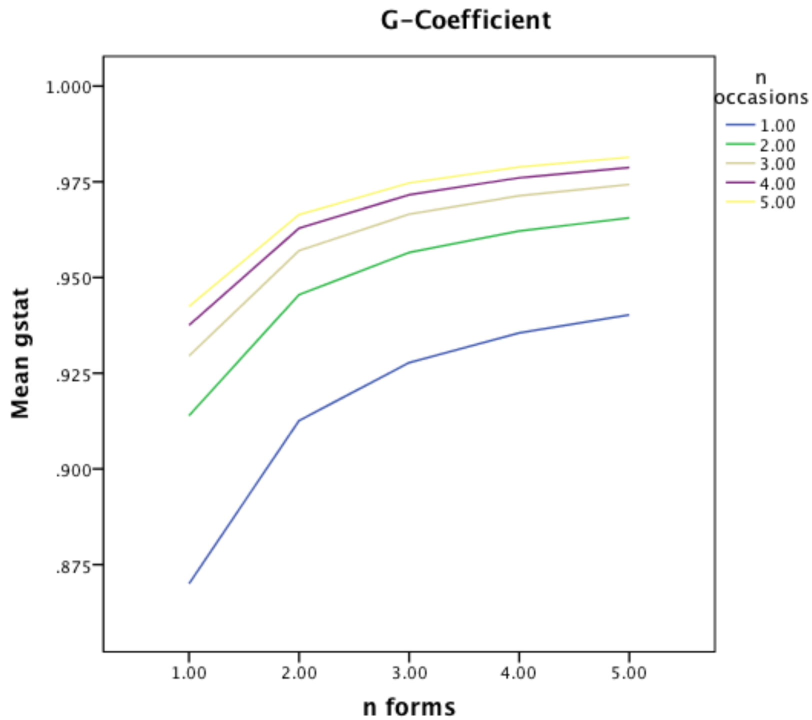
D-Study G Coefficients, Ep^2

n forms	*n* occasions				
	1	2	3	4	5
1	0.87	0.914	0.93	0.938	0.942
2	0.913	0.946	0.957	0.963	0.966
3	0.928	**0.957**	0.967	0.972	0.975
4	0.936	0.962	0.971	0.976	0.979
5	0.94	0.966	0.974	0.979	0.981

Grade 1 WRF: Forms 11, 14 & 15

D-Study Phi Coefficients, Φ

n forms	*n* occasions				
	1	2	3	4	5
1	0.847	0.9	0.919	0.929	0.935
2	0.9	0.938	0.951	0.958	0.962
3	0.919	**0.951**	0.963	0.969	0.972
4	0.929	0.958	0.968	0.974	0.977
5	0.935	0.962	0.972	0.977	0.98



Passage Reading Fluency

Descriptive statistics are presented in Tables 13 and 14. Test-retest reliability results are presented in Table 15. Correlations between each of the 6 forms are presented in Table 16.

Table 13
Descriptive Statistics for Grade 1 Passage Reading Fluency Measures: Session 1

Test Form	*N*	Minimum	Maximum	Mean	Std. Deviation
PRF1.11.1	42	8	204	60.36	41.77
PRF1.13.1	42	4	214	65.05	48.07
PRF1.14.1	19	59	188	107.37	37.13
PRF1.16.1	20	55	178	104.00	37.18

Table 14
Descriptive Statistics for Grade 1 Passage Reading Fluency Measures: Session 2

Test Form	*N*	Minimum	Maximum	Mean	Std. Deviation
PRF1.11.2	41	6	218	64.83	47.03
PRF1.12.2	41	7	191	57.17	43.62
PRF1.13.2	41	8	234	68.56	51.77
PRF1.14.2	38	20	180	89.21	40.29
PRF1.15.2	19	15	150	71.37	38.25
PRF1.16.2	38	12	189	85.05	40.86

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 15 presents results of these analyses. Overall, test-retest reliability was strong, ranging from .83 to .98.

Table 15
Test-retest Reliability of Grade 1 Passage Reading Fluency Measures

Test Form	PRF1.11.2	PRF1.13.2	PRF1.14.2	PRF1.16.2
PRF1.11.1	0.98			
PRF1.13.1		0.98		
PRF1.14.1			0.83	
PRF1.16.1				0.95

Alternate form reliability. Alternate form reliability was evaluated using bivariate correlations among the different forms administered to students. Table 16 displays the results of these analyses. In general, we found strong positive relationships among the alternate forms, with correlations ranging from .93 to .98.

Table 16
Correlation between Alternate Forms of Grade 1 Passage Reading Fluency Measures

Test Form	PRF1.12.2	PRF1.13.2	PRF1.15.2	PRF1.16.2
PRF1.11.2	0.98	0.98		
PRF1.12.2		0.98		
PRF1.14.2			0.96	0.95
PRF1.15.2				0.93

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, 95% and 82% of the variance was associated with the 38 and 13 persons included in the analysis, 0% was associated with forms, and 0% was associated with occasion. The relative error variance was 30.78 for the first analysis and 148.69 for the second, while the absolute variance was 45.16 and 148.69 respectively. The G-Coefficients were .99 for the first analysis and .91 for the second, while the phi coefficients were .87 and .91 respectively.

Passage Reading Fluency: Forms 11 & 13 (teachers 1 & 2)

Grade 1 PRF: Forms 11 & 13
Generalizability ANOVA Table

Facet	*df*	SS	MS	Variance	Proportion
Persons	37	321853.875	8698.753	2143.909	0.953
Forms	1	720.796	720.796	8.429	0.004
Occasions	1	1573.164	1573.164	20.324	0.009
Person*Forms	37	3833.454	103.607	35.611	0.016
Person*Occasion	37	1920.086	51.894	9.755	0.004
Forms*Occasion	1	9.007	9.007	0.000	0.000
Person*Forms*Occasions (Residual)	37	1198.243	32.385	32.385	0.014

Note. Analysis included 38 students, with 2 forms (11 & 13) on 2 occasions.

Error Variances:

Relative, $\sigma^2(\delta_p)$	Absolute, $\sigma^2(\Delta_p)$
30.779	45.155

G-coefficients:

G: Ep^2	Phi: Φ
.986	.979

Grade 1 PRF: Forms 11 & 13

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

n forms	*n* occasions				
	1	2	3	4	5
1	106.503	75.271	64.861	59.655	56.532
2	68.291	**45.155**	37.444	33.588	31.274
3	55.554	35.117	28.304	24.898	22.855
4	49.185	30.097	23.735	20.554	18.645
5	45.364	27.086	20.993	17.947	16.119

Grade 1 PRF: Forms 11 & 13

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

n forms	*n* occasions				
	1	2	3	4	5
1	10.320	8.676	8.054	7.724	7.519
2	8.264	**6.720**	6.119	5.796	5.592
3	7.453	5.926	5.320	4.990	4.781
4	7.013	5.486	4.872	4.534	4.318
5	6.735	5.204	4.582	4.236	4.015

Grade 1 PRF: Forms 11 & 13

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

n forms	*n* occasions				
	1	2	3	4	5
1	77.751	56.681	49.657	46.146	44.039
2	43.753	**30.779**	26.455	24.292	22.995
3	32.42	22.145	18.72	17.008	15.98
4	26.754	17.828	14.853	13.365	12.473
5	23.354	15.238	12.533	11.18	10.369

Grade 1 PRF: Forms 11 & 13

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

n forms	*n* occasions				
	1	2	3	4	5
1	8.818	7.529	7.047	6.793	6.636
2	6.615	**5.548**	5.143	4.929	4.795
3	5.694	4.706	4.327	4.124	3.997
4	5.172	4.222	3.854	3.656	3.532
5	4.833	3.904	3.540	3.344	3.220

Grade 1 PRF: Forms 11 & 13

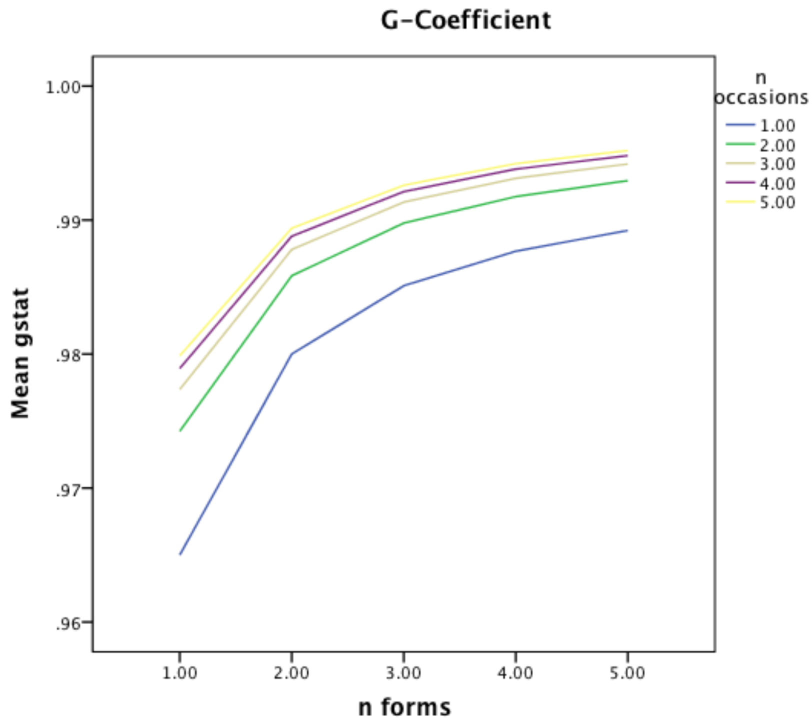
D-Study G Coefficients, Ep^2

n forms	*n* occasions				
	1	2	3	4	5
1	0.965	0.974	0.977	0.979	0.98
2	0.98	**0.986**	0.988	0.989	0.989
3	0.985	0.99	0.991	0.992	0.993
4	0.988	0.992	0.993	0.994	0.994
5	0.989	0.993	0.994	0.995	0.995

Grade 1 PRF: Forms 11 & 13

D-Study Phi Coefficients, Φ

n forms	*n* occasions				
	1	2	3	4	5
1	0.953	0.966	0.971	0.973	0.974
2	0.969	**0.979**	0.983	0.985	0.986
3	0.975	0.984	0.987	0.989	0.989
4	0.978	0.986	0.989	0.991	0.991
5	0.979	0.988	0.99	0.992	0.993


G-Coefficient
Mean gstat
1.00
.99
.98
.97
.96
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 and 16 (teacher 4)

Grade 1 PRF: Forms 14 & 16

Generalizability ANOVA Table

Facet	*df*	SS	MS	Variance	Proportion
Persons	12	75571.5	6297.625	1425.721	0.820
Forms	1	94.231	94.231	0.000	0.000
Occasions	1	105.308	105.308	0.000	0.000
Person*Forms	12	2336.269	194.689	81.321	0.047
Person*Occasion	12	5185.192	432.099	200.026	0.115
Forms*Occasion	1	4.923	4.923	0	0
Person*Forms*Occasions (Residual)	12	384.577	32.048	32.048	0.018

Note. Analysis included 13 students, with 2 forms (14 & 16) on 2 occasions.

Error Variances:

Relative, $\sigma^2(\delta_p)$	Absolute, $\sigma^2(\Delta_p)$
148.685	148.685

G-coefficients:

G: Ep^2	Phi: Φ
.906	.906

Grade 1 PRF: Forms 14 & 16

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

n forms	*n* occasions				
	1	2	3	4	5
1	313.394	197.357	158.678	139.339	127.735
2	256.71	**148.685**	112.677	94.673	83.87
3	237.815	132.461	97.343	79.784	69.249
4	228.368	124.349	89.676	72.34	61.938
5	222.699	119.482	85.076	67.873	57.551

Grade 1 PRF: Forms 14 & 16

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

n forms	*n* occasions				
	1	2	3	4	5
1	17.703	14.048	12.597	11.804	11.302
2	16.022	12.194	10.615	9.730	9.158
3	15.421	11.509	9.866	8.932	8.322
4	15.112	11.151	9.470	8.505	7.870
5	14.923	10.931	9.224	8.239	7.586

Grade 1 PRF: Forms 14 & 16

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

n forms	*n* occasions				
	1	2	3	4	5
1	313.394	197.357	158.678	139.339	127.735
2	256.71	**148.685**	112.677	94.673	83.87
3	237.815	132.461	97.343	79.784	69.249
4	228.368	124.349	89.676	72.34	61.938
5	222.699	119.482	85.076	67.873	57.551

Grade 1 PRF: Forms 14 & 16

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

n forms	*n* occasions				
	1	2	3	4	5
1	17.703	14.048	12.597	11.804	11.302
2	16.022	**12.194**	10.615	9.730	9.158
3	15.421	11.509	9.866	8.932	8.322
4	15.112	11.151	9.470	8.505	7.870
5	14.923	10.931	9.224	8.239	7.586

Grade 1 PRF: Forms 14 & 16

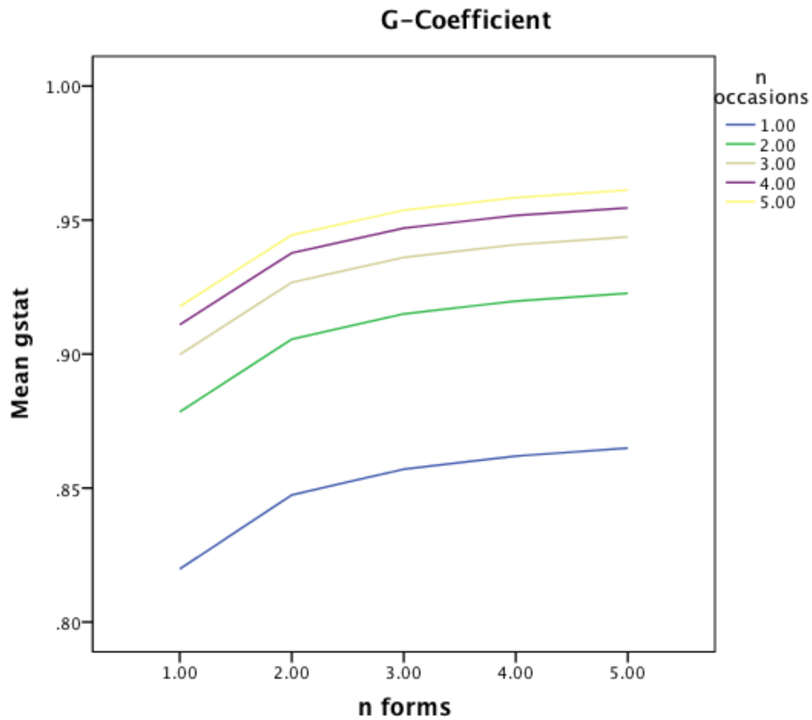
D-Study G Coefficients, Ep^2

n forms	*n* occasions				
	1	2	3	4	5
1	0.82	0.878	0.9	0.911	0.918
2	0.847	**0.906**	0.927	0.938	0.944
3	0.857	0.915	0.936	0.947	0.954
4	0.862	0.92	0.941	0.952	0.958
5	0.865	0.923	0.944	0.955	0.961

Grade 1 PRF: Forms 14 & 16

D-Study Phi Coefficients, Φ

n forms	*n* occasions				
	1	2	3	4	5
1	0.820	0.878	0.900	0.911	0.918
2	0.847	**0.906**	0.927	0.938	0.944
3	0.857	0.915	0.936	0.947	0.954
4	0.862	0.92	0.941	0.952	0.958
5	0.865	0.923	0.944	0.955	0.961



Discussion

Test-retest and alternate form reliability of the four types of grade 1 easyCBM reading assessments were examined in this study. Both test-retest and alternate form reliability of letter sound and phoneme segmenting measures were found to be moderately high. The correlations between measures administered on two testing occasions separated by one week and the correlations between alternate forms of the measures were positive and sufficiently high to suggest the measures' appropriateness for use as progress monitoring tools. Even higher, however, were the test-retest and alternate form reliability estimates of word and passage reading fluency measures. 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 provide additional evidence of the technical adequacy of the grade 1 easyCBM reading measures.

The results of the G- and D-studies were generally mixed. For the G-studies, the majority of variance was routinely attributed to persons, and in some cases overwhelmingly so (e.g., PRF Forms 11 and 13). The results of the analyses with Letter Sounds (LS), Word Reading Fluency (WRF) and Passage Reading Fluency (PRF) measures were generally good. The majority of the variance was routinely associated with persons and the standard errors were reasonably low. Overall, the WRF analyses had the best results, with 85% and 94% of the variance associated with persons in each of the two analyses respectively.

The results of the first analysis for PRF was similarly to the WRF results, with 95% of the variance associated with persons and very low error variances overall, although the results of the second PRF analysis were poorer. Phoneme segmenting (PS) had the poorest results, with a lower amount of variance attributed to persons and a high amount of variance attributed to a person by occasion interaction. We can only speculate as to why PS displayed poorer results, but the person by occasion interaction suggests something changing between testing administrations. The PS measures are perhaps the most difficult to administer of the measures included in this study, and the measure most prone to differences in test administration related to the person administering the tests because unlike the rest of the measures included in this study, the PS measures are administered entirely orally, with the test administrator providing the words to be segmented one at a time. In these measures, differences in the rate at which test administrators provide each word prompt may introduce rater-related sources of error variance. Unfortunately, information about the test administrators was not recorded. Including this information would have made it possible to treat test administrator as an additional source of variance in a three-facet design. Scoring irregularities between occasions is one potential explanation for the poor results. The large person by occasion interaction provides an indication that something of this sort likely occurred.

It is also 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 the number of occasions resulted in a greater increase in dependability than did increasing the number of forms within a single occasion, although often the increase was quite comparable. Unfortunately, how this finding directly connects with practice is unclear given that teachers generally treat each measurement occasion as unique, while the g-theory analyses use the combined information from both testing occasions to produce a single dependability metric. It would be interesting to try different techniques for aggregating the information across two testing occasions to see if the dependability increased by a substantial margin over aggregating information from two test forms within one occasion. When examining the PRF results, however, it is evident that using a single test form on a single occasion is sufficient for dependable measurement and thus no attempt at aggregating information is needed. Using a single form at a single occasion the prophesized g-coefficient ranged from .820 to .965. 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

Full Test form administration order

Teacher	Phoneme Segmenting		Letter Sounds		Word Reading Fluency		Passage Reading Fluency	
	Occasion 1	Occasion 2	Occasion 1	Occasion 2	Occasion 1	Occasion 2	Occasion 1	Occasion 2
1	13 – 12 – 11	13 – 12	11 – 13	12 – 13 – 11	11 – 12	12	11 – 13	13 – 11 – 12
2	11 – 12 – 13	12 – 11	13 – 11	13 – 11 – 12	11 – 12	11 – 12	11 – 13	11 – 12 – 13
3	-	15 – 14	-	15 – 16 – 14	-	12	-	15 – 16 – 14
4	16 – 15 – 14	14 – 16 – 15	16 – 14	14 – 16	14 – 15 – 11	11 – 15 – 14	16 – 14	14 – 16

Test Forms Used for Generalizability Theory Analyses

Teacher	Phoneme Segmenting		Letter Sounds		Word Reading Fluency		Passage Reading Fluency	
	Occasion 1	Occasion 2	Occasion 1	Occasion 2	Occasion 1	Occasion 2	Occasion 1	Occasion 2
1	13 – 12	13 – 12	11 – 13	13 – 11	-	-	11 – 13	13 – 11
2	11 – 12	12 – 11	13 – 11	13 – 11	11 – 12	11 – 12	11 – 13	11 – 13
4	16 – 15 – 14	14 – 16 – 15	16 – 14	14 – 16	14 – 15 – 11	11 – 15 – 14	16 – 14	14 – 16