Best Practices in Oral Reading Fluency Administration

Daniel Anderson, Julie Alonzo, and Gerald Tindal

University of Oregon

Presented at the 2013 NASP Annual Convention

Seattle, Washington

This research was funded by a federal grant awarded to the UO from the United States Department of Education, Institute for Education Sciences: *Reliability and Validity Evidence for Progress Measures in Reading*, R324A100014. The opinions expressed are those of the authors and do not necessarily represent views of the Institute for Education Sciences or the U.S. Department of Education.

For additional information, please contact Daniel Anderson (daniela@uoregon.edu)

Best Practices in Oral Reading Fluency Administration

Educators seeking to evaluate students' reading progress over time often do so through regular administrations of oral reading fluency (ORF) probes. Educational decisions can then be based on both the students' level of performance *and* their corresponding rate of improvement. Instruction can be adjusted according to students' needs based on the observed data. Regular ORF administration provides teachers with a powerful set of data from which decisions can be based. Yet, recommendations in best practices for administration vary by test developer. For example, the developers of easyCBM suggest basing decisions on a single ORF probe (Alonzo & Tindal, 2012), while the DIBELS*Next* developers suggest administering three probes in succession, taking the median score (Dibels*Next*, 2011). There are two primary reasons for a median-score approach over a single probe approach: (a) to help alleviate test form effects, and (b) to produce more stable estimates of students' flueny (see Petscher & Kim, 2011). Yet, if sufficient passage comparability can be established across test forms, with single testing administrations demonstrating reasonably high reliability, then a single passage administration approach may be more resource efficient.

The purpose of this paper is to explore the reliability of ORF scores under various testing conditions, utilizing Generalizability and Decision Studies (G- and D-studies; Brennan, 2001). The G-study allows one to examine the variance associated with multiple facets of the measurment process (e.g., test forms, testing occasion, etc.) while the D-study allows one to examine how the reliability of the measurement would change with different applications of the facets. The D-study, therefore, allows exploration of how ORF reliability would change when one versus multiple passages are administered, as well as with one versus multiple testing

occassions. Determination of best administration practice, however, depends first upon the reliability and comparability of the test forms.

ORF Passage Comparability and Measurement Error

There are two primary sources of measurement error in ORF probes – those due to the individual test form and those due to differences in comparability across forms (see Christ & Ardoin, 2009; Christ & Silberglitt, 2007; Francis et al., 2008). Because ORF probes are used to evaluate student progres – that is, score *changes* over time in addition to score *level* at any single point in time – measurement error concerns are compounded. Test forms of insufficient comparability, or with overly large standard errors, are likely to result in inferences of student achievement that are artifacts of the testing process rather than accurate representations of student skill.

ORF Passage Comparability. Evaluating score changes compounds measurement error, and unreliability in test forms leads to a greater unreliability in the score changes (Bereiter, 1963). Although this concern can be mitigated with multiple measurement occassions (i.e., more than two; Singer & Willett, 2003) and more advanced statistical techniques (e.g., multilevel or latent growth curve models; see Preacher, Wichman, Briggs, & MacCallum, 2008; Raudenbush & Bryk, 2002), statistics alone cannot overcome deficiencies in measurement.

The comparability of scores from one testing occasion to the next is critical to the validity of ORF score-based inferences. Using a median-score administration method, one may minimize the chances that an observed score is the result of an unusually easy or difficult test form. For example, imagine a situation with three ORF probes: one of high difficulty and two of average difficulty. If a single administration practice was used, then students' scores would likely change from one testing occasion to the next based on the occasion in which they received the difficult

form, and not necessarily because of any specific change in reading ability. Or, similarly, a substantial change in ability may have occured, but not be reflected in the scores (i.e., flat growth). Using the median score approach, all three test forms would be administered during a single testing occasion and the score the student received on the difficult form would likely be dropped (i.e., it would not represent the median score). In practice, form effects such as these have been observed. Christ and Ardoin (2009), for example, found average differences as large as 46 words read correctly per minute between the easiest and most difficult test forms, while Francis et al. (2008) found that the order in which students received test forms changed the rate and shape of the estimated growth curve.

Of course, the degree to which dropping scores is of benefit depends directly on the comparability of the ORF probes. Francis et al. (2008) contend that ORF probes should be equated to ensure comparability, and that high raw-score correlations between forms does not guarantee equivalence. If an equating process has been conducted during measurement development, however, and the probes *are* of sufficient comparability, then one of the primary reasons for taking a median-score administration approach is no longer of concern. Yet, to date, no commercial test developers of oral reading fluency measures report scores on an equated scale score. Thus, if Francis et al. are correct, only the median-score administration approach would be feasible.

But, could it also be possible that careful instrument development and administration procedures could lead to *adequately* comparable forms? Whether the metric of the ORF is reported on a raw score or an equated scale score, if adequate comparability of forms can be established, then the primary reason for taking a median-score approach becomes concerns with measurement error inherent in any one form, but not across forms. That is, the argument

becomes that a single, one-minute timed fluency probe is not sufficient to adequately assess the fluency of a student.

ORF Standard Errors. A considerable amount of research has investigated the reliability of ORF passages (see Deno et al., 1982; Fuchs, 2004; Hintze, Daly, & Shapiro, 1998), but comparatively little has investigated the standard error of measurement (SEM) for ORF passages. Poncy, Skinner, and Axtell (2005) applied Generalizability Theory and Decision studies to estimate the SEM under different testing conditions. The authors estimated the SEM of a single passage to be 12 or 18 words, depending on on the construction of the passage. This estimate was reduced to 5 or 7 words when 3 passages were used. The practical implications of the differences in SEM can perhaps best be evaluated through the construction of a confidence interval. For example, imagine a student who is administered a single ORF passage and receives a score of 100 words read correctly per minute. Using the results obtained by Poncy et al., and assuming the more careful test construction was used to reduce the size of the SEM, then we could be 95% confident that the student's "true score" would range between $100 \pm (2 * 12) = 76$ and 124 when a single passage was administered. If we administered three passages, then the range of the confidence interval would be reduced to $100 \pm (2 * 5) = 90$ to 110.

Christ and Silberglitt (2007) used a large sample across multiple years in grades 1-5 to estimate the average SEM when a median-score administration was used. The authors found that the variance in passages universally increased with grade level, leading to higher SEMs.The authors explored how the SEM may change based on relatively homogenous, typical, or heterogeneous student populations, as well as by the estimated reliability of the form. The median SEM across grades was 10 words per minute– slightly higher than the SEM estimated by Poncy et al. (2005), and actually near the lower SEM bound for a single passage. Not

surprisingly, Christ and Silberglitt also found that the SEM decreased as reliability increased. Overall, the estimated SEM ranged from 4 to 15 words read correctly per minute.

The size of the SEM determines the level of confidence one can have in the estimates of students' reading fluency. The key question, however, is whether additional educational resources should be dedicated to the testing process to obtain smaller SEMs. For an individual student, the additional resources may be quite minimal (i.e., two additional test forms and a few extra minutes of testing). However, during interim benchmark testing, as within a schoolwide response to intervention framework (see Curtis, Sullivan, Alonzo, & Tindal, 2011) the additional resources may be quite substantial.

Practical Repercussion in Administration Choice

The results of Poncy et al. (2005) and (Christ & Silberglitt, 2007) are reasonably consistent. However, the evidence for best practice in ORF administration remains quite unclear. For example, Poncy et al. suggest that if careful construction is used, the SEM for a single passage is around 12 words per minute. Is this too large? If so, then the immediate conclusion may be that a single passage administration practice is inappropriate. Yet, if a move toward a multiple passage administration practice were taken, how much of a reduction in the SEM would be observed? Similarly, would the practical consequences of such a move (i.e., increased school resources devoted to testing) be worth the drop in measurment error? And how would the increased resources directed at one construct of reading (fluency) affect the resources available for assessing other contructs of reading (i.e., comprehension and vocabulary)?

When considering the best administration practice for oral reading fluency probes, one must consider both the precision of the estimates and the practical consequences of one approach over another. Although a median-score approach will nearly always have psychometric benefits,

such benefits must be considered in light of the practical repercussions. If sufficiently useful scores can be obtained from a single passage, then administering only one passage may be a worthwile option for districts, particularly those in which resources are tight. The purpose of this study is to explore how the reliability and standard errors of ORF passages change under various administration designs in grades 1-5. Similar to Poncy et al. (2005), we use G- and D- studies to estimate reliability and standard errors under different conditions of measurement. In our study, all scores were obtained using a single passage approach.

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 administered a battery of easyCBM assessments. Data were gathered on two separate occasions, one week apart. Each day, students were administered a series of alternate grade-appropriate ORF passages in one-on-one settings. Assessments were generally counter-balanced to control for order effects.

The data used for this study were part of a larger study investigating the alternatate form and test-retest reliability of the easyCBM assessments in reading (in addition to the ORF data analyzed in this study, data were gathered on early literacy measures in grades K-2). Thus, the order in which the different test forms were administered was not always consistent for a fully crossed design. Rather, the test forms were nested within students. Nested designs lead to fewer sources of error variance being estimatable (despite still being present). For example, the form the student was administered is one possible source of error (e.g., the form could be more or less difficult than other forms). Yet, with a design where the forms are nested within persons (F:P), the variance uniquely attributable to the form is indistinguishable from the variance attributable

to a form-by-person interaction. Given that we were interested in the variance uniquely attributable to test forms, we restricted the full sample to include only a subsample of students who were administered the test forms in a fully crossed design.

Generally, there were multiple possible (fully crossed) analyses within each grade. In this manuscript, we present only one analysis from within each of grades 1-5. Each analysis presented here was selected based on its representativeness of the other analyses conducted within the grade. For a more complete description of all analyses, including those not presented here, we refer readers to Alonzo, Lai, Anderson, Park, and Tindal (2012); Anderson, Lai, Park, Alonzo, and Tindal (2012); Anderson, Park, Lai, Alonzo, and Tindal (2012); Lai, Park, Anderson, Alonzo, and Tindal (2012); and Park, Anderson, Alonzo, Lai, and Tindal (2012). The number of test forms administered to students during each occasion varied by grade. Table 1 reports the sample size and testing procedures for each grade for the analyses reported here. Note that the test forms were generally, but not always, counterbalanced between occasions to control for order effects.

Measures

For this study, we used the easyCBM passage reading fluency (PRF) measures, standardized measures of Oral Reading Fluency. The easyCBM PRF measures were developed in 2006 by researchers at the University of Oregon (Alonzo, Tindal, Ulmer, & Glasgow), using an iteritive process of instrument development, review for issues related to potential bias and grade-level-appropriateness, revision, field testing, analysis of form comparability, revision, additional field testing, and finally calculation of normative performance. Passages were written following word count and grade-level guidelines. Data on passage difficulty, such as the Flesch-

Kincaid readability estimates, were used to bring the passages into closer alignment during initial writing and revision.

During the field testing phase of instrument development, passage comparability was analyzed using correlations and ANOVA to test mean differences between the different forms of the measures. Through this process, passages that were most similar in difficulty were identified, and the difficulty level of the remaining passages was either increased or decreased, to bring them into closer alignment in terms of difficulty. During instrument development and field testing, all 20 alternate forms at each grade level were administered to the same group of students over the course of one week. Thus, each student served as his/her own control to reduce the confound that would be introduced if the samples had varied by test form (see Alonzo & Tindal, 2007, for a more detailed documentation of instrument development, and Jamgochian et al, 2010; Sáez et al., 2010, and Lai et al., 2010, for more detailed documentation related to validity studies).

Trained researchers administered grade-level PRF passages using standardized test administration procedures. Tests were administered individually to students in a one-on-one setting. Researchers read from printed test administration instructions on the administrator copy of the PRF measures while students were provided a student copy of the measure. As students read aloud from their copy of the PRF measure, test administrators marked any words read incorrectly or skipped, indicating self-corrections (not counted as errors) if students made any. At the end of 60 seconds, test administrators marked the last word students had read and calculated the total correct words read per minute by subtracting the total number of errors or skipped words from the total words read. If students hesitated for more than three seconds while

reading, test administrators provided them with the word and asked them to continue, marking the word the student was provided as an error.

Analyses

For our generalizability theory study (G-Study) we calculated the variances associated with persons and two facets: forms and occasions. Figure 1 provides a hueristic for the estimable G-Study variance components. There are three overlapping circles in the figure, each representing a source of variance – persons (i.e., the object of meausrement) and two error variances, forms and occassions. Note that, because the design was fully crossed we could estimate each source of variance uniquely, as well as all interactions among the variance components (e.g., persons by forms). Our G-Studies were then followed up with decision studies (D-Studies) to help determine the necessary conditions for reliable measurement. For example, to obtain reliable estimates of students' ability, such that decisions could be made with reasonable confidence, should students be administered 1, 2, 3, 4, or 5 forms during any one occasion? Similarly, does increasing the number of testing occasions increase the reliability of the estimate, and at what point is a reliable estimate obtained? Although typical CBM administration recommendations only discuss adminsitration within a single occasion, we include *occasion* as a relevant facet in determining reliability. The G-study provides information on the sources of error in the measurement process while the D-study provides information on potential ways that the measurement process could be changed to produce more reliable results. Brennan (2001) recommends phi coefficients and absolute error variances be interpreted when the results of the test – rather than normative comparisons with peers – will be used as the basis to form decisions.

Results

We present our results from an overall perspective. That is, we emphasize overall trends across grades, rather than specific results of any one grade. As previously mentioned, the results reported here are only a portion of the entire results, as reported in Alonzo et al. (2012); Anderson, Lai, et al. (2012); Anderson, Park, et al. (2012); Lai et al. (2012); and Park et al. (2012). Table 2 presents the variance components for each facet of measurement across grades, obtained from the G-Study, while Table 3 presents the estimated standard errors and dependability coefficients for different administration practices (one to five forms within one or two testing occasions). The dependability coefficients are also plotted in Figure 2 for a sample of two grades. The figure illustrates the change in dependability coefficient by the corresponding levels of the facets. Note that the figure plots the dependability coefficients for more potential occasions (5) than are reported in the Table 3.

As reported in Table 2, approximately 79-95% of the total variance was attributable to persons – the object of measurement. Forms were a relatively neglible source of error variance, with essentially no variance attributable to the test forms in grades 1, 2, and 4, and only 2% or 3% of the variance attributable to test forms in grades 3 and 5 respectively. The person by form interaction was a larger source of error variance overall, ranging from 0-4%, suggesting that, while the form was not a substantial source of error variance overall, it may have played a larger role for any individual student. That is, unique characteristics of the student (e.g., topical areas of interest) interacted with test forms to produce, generally, larger error variance overall than the test form did on its own. Occasion was generally a larger source of error variance than test forms, but was again quite neglible, ranging from 1-4% across grades. For grade 3, there was a sizeable person by occasion interaction (6% of the total variance). For all other grades, however, the person by occasion interaction was minimal (0-2% of the total variance). Given the small sample

size, (n = 28 for grade 3) it is likely that the size of the interaction was sample specific, and may have been the result of a unique occurance. For example, perhaps these students had a substitute teacher during one of the testing occasions. Across all grades, there was essentially no test form by occasion interaction. Finally, approximately 1-14% of the total variance was unattributable to any unique facet, or interaction among facets.

All the preceding numbers reported, and displayed in Table 2, were estimated for the conditions of measurment used in the study. That is, referring to Table 1, we can see that students in grade 2 were administered three test forms (11, 12, and 13) on two occasions. Under this condition of measurement, approximately 88% of the total variance was attributable to persons. Under different conditions of measurement, the variances may be redistributed. Examining Table 3, we can see the absolute standard error and dependability coefficients for various conditions of measurement, including the observed condition used in the study. For grade 2 the absolute standard error for the condition of measurement used in the study would be 7.07, while the absolute dependability coefficient would be .96. However, we can also use the data to predict what these values would have been had only one form been administered on one occasion, or any of a number of other possible measurement conditions.

Across grades, the absolute standard errors for one test form on one occasion ranged from 10.32 to 16.23. As previously mentioned, we can use these standard errors to produce confidence intervals for any one student's score. So, for a student at grade 1 who reads 75 words correct, we can be 95% confident that his or her true score lies in the range $75 \pm (10.32*1.96) \approx 55$ to 95. Using this same formula, we can see how our our range increases at grade 3, where the standard error is 16.23. At grade 3, we would be 95% confident that a students true score would lie roughly in the range of 43 to 107. Increasing the number of test forms from 1 to 3 universally

reduced the size of these standard errors. The reduction was generally in the range of 3-4 words. Increasing the number of occasions from one to two also reduced the standard errors. When only one test form was administered during each occasion, increasing the number of occasions from one to two decreased the standard errors by approximately 2-3 words.

The absolute dependability coefficients (phi) for one test form on one occasion ranged from .79 at grade 5 to .95 at grade 1. These coefficients can be interpreted similarly to Cronbach's alpha coefficient for reliability. The size of the coefficients, again, universally increased when three test forms were administered in comparison to one, ranging from .88 at grade 3 to .98 at grade 1. Generally, increasing the number of test forms resulted in roughly the same increase in the dependability coefficient as increasing the number of testing occasions. Further, the results were quite similar across grades. Figure 2 displays the relation between the phi coefficents and multiple levels of each facet. As can be seen, the results are quite similar between the two presented grades, which are quite representative of the other grades (see Alonzo et al. 2012; Anderson, Lai et al., 2012; Anderson, Park et al., 2012; Lai et al., 2012; & Park et al., 2012).

Discussion

The best ORF administration practice may well depend on the context in which the measure is being administered. The decision whether to use a median-score or single probe depends upon both the comparability of ORF probes *and* the measurement error inherent in any one form. Although both conditions of measurement are important for either administration practice, use of a single probe demands that alternate forms of the measure are comparable to one another. A median-score approach is more general in its assumptions, in effect 'washing out' some of the between-form differences. For a single test form administration practice, the

importance of sufficient comparability across forms becomes magnified. If the comparability of the psychometric properties (e.g., reliability, difficulty, etc.) cannot be established across multiple test forms (Francis et al., 2008), then the single-form administration practice may be infeasible.

However, the results of the current study suggest that administering a single form of a tightly-constructed oral reading fluency passage results in sufficiently dependable measurement, with an estimated dependability coefficient (phi) within the recommended range of reliability (Ponterotto & Ruckdeschel, 2007). This finding underscores the importance of addressing comparability of forms in the process of constructing oral reading fluency measures. This goal might be attained through scaling ORF scores, thus smoothing out form differences using a psychometric approach (Crist & Ardoin, 2009). As in our current study, it might also be attained through measurement development efforts whereby text characteristics (genre, vocabulary, sentence length, syntax, and story structure) are controlled to reduce cross-form variation (Alonzo & Tindal, 2007).

In our current study, we found that administering one additional form of oral reading fluency measure reduced the estimated absolute standard error by approximately 3 words per minute. The impact of this reduction of absolute standard error, of course, varies slightly by grade. In first grade, for instance, expected weekly growth on easyCBM® PRF measures for students in the 50th percentile is .87 words per minute; by fifth grade, expected weekly growth for students at the 50th percentile is .63 words per minute (easyCBM.com). If schools are following the testing schedule recommended by the authors of easyCBM®, where PRF measures are administered no more frequently than every other week, reducing the absolute standard error of measure by three words per minute may have minimal impact on the overall interpretability of

the scores. In most settings it is likely that administering a single form of tightly-constructed ORF measures during each assessment cycle would be sufficient (e.g., for interim benchmarking and progress monitoring), particularly when weighed against the costs associated with a median-score approach.

The difference in testing time between a single passage versus a median-score approach can be quite substantial when aggregated across a large district. For example, suppose a school district had 10,000 students across grades 1-5. Generally, districts of this size hire and train paraprofessionals to administer the ORF probes to all students to minimize the impact on teachers and reduce the potential for inter-rater discrepancies. Under a single-probe adminsitration approach, enough paraprofessionals would need to be hired, trained, and paid to cover a minimum of 10,000 minutes – or approximately 167 hours – of testing (not accounting for transition times, etc.). However, under a median-score approach a minimum of 30,000 minutes – or approximately 500 hours – of testing would be required. Further complicating matters, under a typical Response to Intervention plan, district-wide screening occurs three times a year (fall, winter, and spring). Thus, in our hypothetical district, the annual difference between a single or median-score administration practice, would be the difference between roughly 500 and 1,500 total hours of testing across grades 1-5. Assuming a paraprofessional staff receiving minimum wage, this is a difference of approximately $9000 in direct cost to the district. In districts where district-wide screening assessments are administered by certified staff or where those hired to administer the assessments receive more than minimum wage, the overall cost for the district would be substantially more with a median, versus single passage, approach.

However, there may be times when a median-score approach would still be recommended. For instance, if the assessments being used are known to vary in difficulty, or

shift between genre from form to form, administering three different forms and using the median score would be preferable. There may also be times when broadening the assessment data being considered beyond oral reading fluency is important. When the decision to be made is high stakes (e.g., referral to special programs, exiting from a particular tier of instructional supports), a different approach may well make more sense. For instance, rather than relying on ORF performance alone, a combination of performance on a variety of measures sampling from the larger construct of reading (including, for instance, reading comprehension and vocabulary in addition to oral reading fluency) may be called for. One question that we did not address in this reasearch, but which remains a fruitful area for future consideration, is the degree to which the reliability and validity of interpretations of student reading skill and growth in performance might be enhanced by widening the lens through which we measure reading. A reduction in the expense (in time, materials, and human resources) associated with administering multiple alternate forms of a single measure type (such as ORF), might enable districts to sample more broadly from the construct of reading, and thus increase the reliability and validity of decisions related to student reading performance.

References

Alonzo, J., Lai, C. F., Anderson, D., Park, B. J., & Tindal, G. (2012). An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM reading assessments: Grade 4 (Technical Report No. 1219). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2012). *Teachers' manual for regular easyCBM: Getting the most out of the system.* Retrieved June 6, 2012, from http://easycbm.com/teachers/auth/index.php

Anderson, D., Lai, C. F., Park, B. J., Alonzo, J., & Tindal, G. (2012). An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM reading assessments: Grade 2 (Technical Report No. 1217). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Park, B. J., Lai, C., F., , Alonzo, J., & Tindal, G. (2012). An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM reading assessments: Grade 1 (Technical Report No. 2016)

. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Bereiter, C. (1963). Some persisting dilemmas in measurement of change. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press.

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

Christ, T. J., & Ardoin, S. P. (2009). Curriculum-based measurement of oral reading: Passage equivalence and probe-set development. *Journal of School Psychology, 47*, 55-75. doi: 10.1016/j.jsp.2008.09.004

Christ, T. J., & Silberglitt, B. (2007). Estimates of the standard error of measurement for curriculum-based measures of oral reading fluency. *School Psychology Review, 36*, 130-146.

Curtis, Y., Sullivan, L., Alonzo, J., & Tindal, G. (2011). Context and process for implementing RTI. In E. S. Shapiro, N. P. Zigmond, T. Wallace & D. Marston (Eds.), *Models for implementing Response to Intervention*. New York: Guilford.

Deno, S., Marston, D., Mirkin, P., Lowry, L., Sindelar, P., & Jenkins, J. R. (1982). The use of standard tasks to measure achievement in reading, spelling, and written expression: A normative and developmental study. Minneapolis, MN: Institute for Research on Learning Disabilities.

Dibels*Next*. (2011). Dibels Oral Reading Fluency Retrieved February 14, 2012, from https://http://www.mclasshome.com/wgenhelp/dnext/DIBELS_Next/Assessment_and_Scoring/DORF_Details.htm

Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. R. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology, 46*, 315-342. doi: 10.1016/j.jsp.2007.06.003

Fuchs, L. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*, 188-192.

Hintze, J. M., Daly, E. J., & Shapiro, E. S. (1998). An investigation of the effects of passage difficulty level on outcomes of oral reading fluency progress monitoring. *School Psychology Review, 27*, 433- 445.

Lai, C. F., Park, B. J., Anderson, D., Alonzo, J., & Tindal, G. (2012). An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM reading

assessments: Grade 5 (Technical Report No. 1220). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Park, B. J., Anderson, D., Alonzo, J., Lai, C. F., & Tindal, G. (2012). An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM reading assessments: Grade 3 (Technical Report No. 1218). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Petscher, Y., & Kim, Y.-S. (2011). The utility and accuracy of oral reading fluency score types in predicting reading comprehension. *Journal of School Psychology, 49*, 107-129. doi: 10.1016/j.jsp.2010.09.004

Poncy, B. C., Skinner, C. H., & Axtell, P. K. (2005). An investigation of the reliability and standard error of measurment of words read correctly per minute using curriculum-based measurement. *Journal of Psychoeducational Assessment, 23*, 326-338. doi: 10.1177/073428290502300403

Ponterotto, J. G., & Ruckdeschel, D. E. (2007). An overview of coefficient alpha and a reliability matrix for estimating adequacy of internal consistency coefficients with psychological research measures. *Perceptual and Motor Skills, 105*, 997-1014. doi: 10.2466/pms.105.3.997-1014

Preacher, K. J., Wichman, A. L., Briggs, N. E., & MacCallum, R. C. (2008). *Latent Growth Curve Modeling*. Thousand Oaks, CA: Sage.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Second ed.). Thousand Oaks, CA: Sage.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurence*. New York, NY: Oxford University Press.

Table 1

Sample administration by grade

Grade	Total *n*	Condition	Test forms: Day 1	Test forms: Day 2
1	38	1	11-13	13-11
		2	11-13	11-13
2	31	1	13-12-11	13-11-12
		2	11-12-13	12-13-11
3	28	1	16-15-14	16-14-15
		2	14-15-16	15-16-14
4	39	1	13-12-11	13-11-12
		2	11-12-13	12-13-11
5	13	1	8-9-10-12	9-10-8-12

Note. For each grade, roughly half the sample was assigned to each condition. Data were combined across conditions for all analyses.

Table 2

Variance Components for G-Theory Analyses

Grade	Persons	Forms	Occasion	Persons*Forms	Persons*Occasion	Forms*Occasion	Residual
1	2143.91 (.95)	8.43 (0.00)	20.32 (.01)	35.61 (.02)	9.76 (.00)	0.00 (.00)	32.39 (.01)
2	1306.29 (.88)	5.87 (.00)	16.44 (.01)	26.17 (.02)	29.24 (.02)	4.98 (.00)	94.12 (.06)
3	1237.18 (.82)	21.83 (.02)	36.58 (.02)	56.67 (.04)	83.81 (.06)	3.52 (.00)	61.07 (.04)
4	1363.10 (.88)	0.00 (.00)	65.52 (.04)	31.15 (.02)	15.25 (.01)	7.90 (.01)	71.91 (.05)
5	621.75 (.79)	26.74 (.03)	18.96 (.02)	0.00 (.00)	9.46 (.01)	0.00 (.00)	108.55 (.14)

Note. Proportion displayed in parentheses. Residual term represents a person by form by occasion interaction.

Table 3

Predicted absolute standard errors and dependability (reliability) coefficients by administration practice

Reliability index	Grade	D studies									
		n Occasions	1	1	1	1	1	2	2	2	2	2
		n Forms	1	2	3	4	5	1	2	3	4	5
SE, $\sigma(\Delta_p)$	1	-	10.32	8.26	7.45	7.01	6.74	8.68	6.72	5.93	5.49	5.20
	2	-	13.30	10.55	9.46	8.86	8.48	10.22	7.98	7.07	6.58	6.26
	3	-	16.23	13.85	12.96	12.50	12.21	13.08	10.75	9.86	9.38	9.08
	4	-	13.85	11.67	10.85	10.42	10.15	10.56	8.71	8.00	7.63	7.39
	5	-	12.80	9.80	8.57	7.89	7.45	9.76	7.40	6.42	5.87	5.52
Phi, Φ	1	-	.95	.97	.98	.98	.98	.97	.98	.98	.99	.99
	2	-	.88	.92	.94	.94	.95	.93	.95	.96	.97	.97
	3	-	.82	.87	.88	.89	.89	.88	.92	.93	.93	.94
	4	-	.88	.91	.92	.93	.93	.92	.95	.96	.96	.96
	5	-	.79	.87	.89	.91	.92	.87	.92	.94	.95	.95

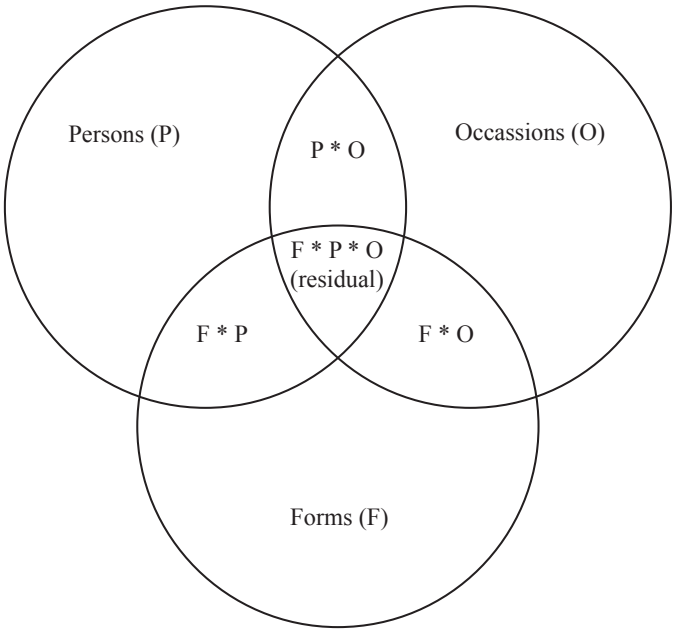
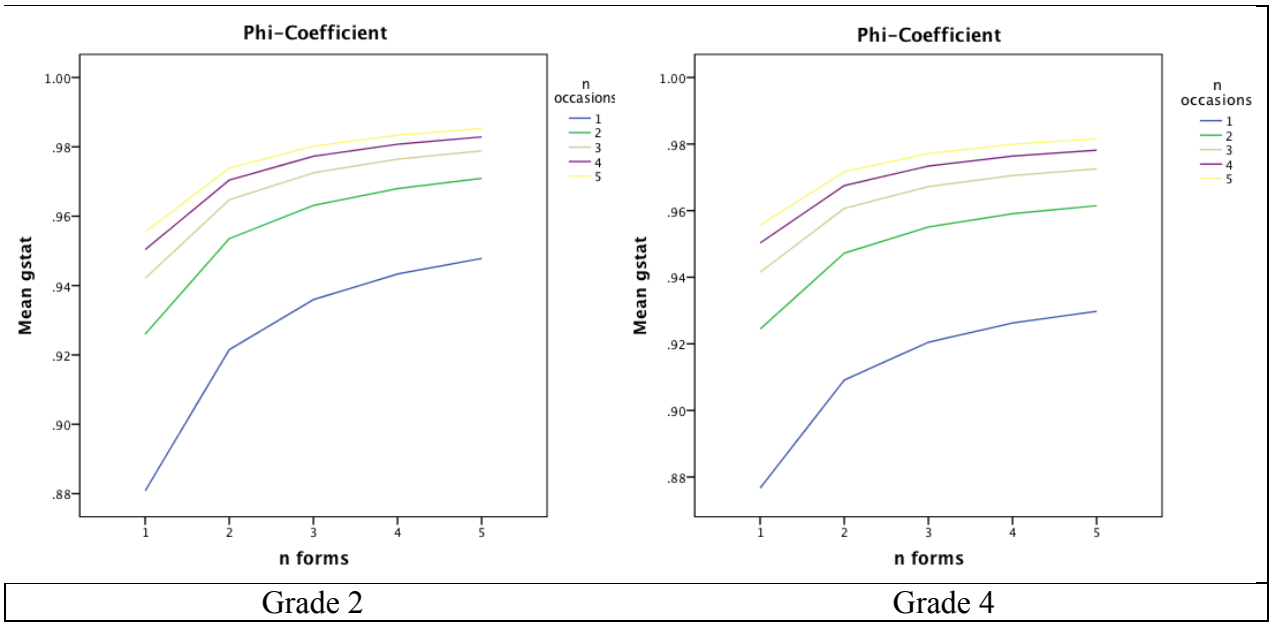


Figure 1. Hueristic of variance components estimated for all G-Studies. Each circle in the figure represents a separate variance component estimated in the model, with the overlapping portions representing interactions.



Grade 2 Grade 4

Figure 2. Absolute dependability coefficients for a sample of two grades. Figure displays how the dependability coefficients were predicted to change based on various conditions of measurement. Each line represents a different number of testing occassions.