The Impact of Probe Difficulty Variation on Brief Experimental Analysis of Reading Skills

Sterett H. Mercer

University of British Columbia

Lauren Lestremau Harpole, Rachel R. Mitchell, Chandler McLemore, and Christina Hardy

The University of Southern Mississippi

Author Note

Sterett H. Mercer, Department of Educational & Counselling Psychology and Special Education, University of British Columbia; Lauren Lestremau Harpole, Rachel R. Mitchell, Chandler McLemore, and Christina Hardy, Department of Psychology, The University of Southern Mississippi.

This research was supported, in part, by a grant from the Early Career Awards Program of the Society for the Study of School Psychology.

Correspondence concerning this article should be addressed to Sterett H. Mercer, Department of Educational & Counselling Psychology and Special Education, 2125 Main Mall, Vancouver, BC V6T 1Z4, Canada. E-mail: sterett.mercer@ubc.ca

Abstract

The impact of variation in probe difficulty on the ability to replicate results in brief experimental analysis (BEA) of reading was examined. In the first phase of the study, 41 first- and second-grade students completed 16 reading probes. Calculations of probe difficulty were used to identify Low and High Variability probe sets. In the second phase of the study, 40 second- through fifth-grade students' performance on two reading interventions was compared in a BEA-like task. The best-performing intervention was unlikely to be replicated on either probe set (i.e., for only 43% of students); rather, the best determinant of intervention replication was each students' average difference in performance across the two interventions. The best-performing intervention was more likely to be replicated (i.e., 60% of students) when averages of two trials per intervention were compared. These results are discussed in the context of developing rules for determining the best-performing intervention in academic BEA.

Keywords:

reading, brief experimental analysis, curriculum-based measurement, generalizability theory

The Impact of Probe Difficulty Variation on Brief Experimental Analysis of Reading Skills

School psychologists increasingly are encouraged to implement evidence-based interventions as part of their practice (Kratochwill & Stoiber, 2002), and numerous evidence-based interventions to improve academic skills have been developed. Despite research supporting the efficacy of specific academic interventions, there is no guarantee that comparable effects will be found when these evidence-based interventions are implemented with a particular student. Logically, the likelihood of intervention success may be improved by selecting an intervention with the largest reported effect size, yet this strategy will not ensure that the best and most efficient intervention will be selected for a particular student. These concerns regarding intervention effectiveness for individual students can be partially addressed by conducting ongoing progress monitoring and formative evaluation (Fuchs & Fuchs, 1986), and brief experimental analysis (BEA) procedures can be used as part of this process to evaluate the differential effectiveness of academic interventions with individual children (e.g., Daly, Persampieri, McCurdy, & Gortmaker, 2005).

In academic BEA, performance on specific interventions relative to each other and baseline is compared in a multi-element design. Academic BEA procedures vary across studies; however, a series of academic interventions that vary in terms of intrusiveness or complexity are typically administered in alternating or random sequence, with the best-performing intervention confirmed in a mini-withdrawal (Jones, Wickstrom, & Daly, 2008). For example, in Jones et al. (2008) correct words per minute (CWPM) and errors per minute (EPM) on grade-

level reading passages were compared under each of the following conditions: baseline, incentive (i.e., a reward was given for meeting goal), repeated reading (i.e., the student read the passage three times), listening passage preview with phrase drill (i.e., the examiner read the passage first while the student followed along, and errors that were made during the student's initial reading were corrected and rehearsed), and easier material (i.e., the student read a passage one grade level lower than baseline). One trial of each intervention was tested in order of increasing intervention complexity, and the best-performing intervention on the initial trials (repeated reading) was confirmed in an additional trial. Because the interventions can be administered in a time-efficient manner and performance can be compared on specific probes immediately following intervention trials, academic BEA has been promoted as a time-efficient method to empirically inform intervention selection for individual students as part of evidence-based practice (Jones et al., 2009).

Reliability of Curriculum-Based Measures

Although BEA methods have been adapted and applied to diverse academic skills such as reading fluency (Daly, Bonfiglio, Mattson, Persampieri, & Foreman-Ya, 2006), letter sound fluency (Petursdottir et al., 2009), math computation (VanDerHeyden & Burns, 2009), and basic writing skills (Burns, Ganuza, & London, 2009), critical issues need to be addressed prior to more widespread use of BEA methods. Namely, most academic BEA methods rely on curriculum-based measures (CBM) of academic skills, and CBM has questionable reliability under some circumstances. Several studies (Christ & Ardoin, 2009; Hintze, Owen, Shapiro, & Daly, 2000; Poncy, Skinner, & Axtell, 2005) have found that the reliability of reading CBM scores is lower when used for absolute decisions (in reference to a specific score) as compared to

relative decisions (rank-ordering students). The importance of this finding was stressed by Christ and Ardoin (2009) in their comment that "the field of school psychology has a dependent variable problem at the same time that a paradigm shift is underway" (p. 72). Specifically, reading CBM can exhibit questionable reliability when scores are compared to an absolute standard of performance, such as scores on a prior occasion or specific cut-scores, which is how CBM is often used when determining a student's response to a particular intervention.

The extent to which absolute decisions are limited by the reliability of reading CBM is a function of two factors: the quality of the probe set and the number of probes administered. In Poncy et al. (2005), 37 third-grade students read 20 different CBM probes (Good & Kaminski, 2002). Results indicated that with one probe administered, reliability for absolute decisions was .81, with a standard error of measurement (SEM) of approximately 18 CWPM. Based on the study's results, administering and averaging performance on three probes would considerably improve reliability for absolute decisions (.93); however, the SEM was still large (10 CWPM). Both reliability and SEM can be improved by additional field testing to select optimal reading probes. For example, in Poncy et al. (2005), reliability for absolute decisions based on single probes improved to .89 and SEM was reduced to 12 CWPM when probes were excluded from the set based on average scores across students greater than 5 CWPM away from the overall average for the probe set. In addition, more recently applied methods to select optimal probes have resulted in greater improvements to reliability and SEM (e.g., Christ & Ardoin, 2009).

Impact of SEM on BEA Decisions

Comparisons of absolute values derived from a limited number of observations are central to BEA methodology, raising concerns about the reliability of decisions resulting from

this process. In academic BEA, performance on CBM probes following the brief implementation of various interventions is compared by intervention and relative to baseline to determine the most effective intervention strategy for a particular student. The magnitude of SEM for absolute decisions is particularly relevant to BEA decisions given that SEM characterizes the expected variation around CBM scores. In both Poncy et al. (2005) and Christ and Ardoin (2009), SEMs for individual probes remained at 10-12 CWPM following application of methods to improve probe sets. Based on this SEM, in 68% of future probe administrations, one would expect scores to vary by 20-24 CWPM without any change in reading skill. This range of expected variation is reduced when considering averages of two probes, and averages of two probes are available for comparison so long as mini-withdrawals are conducted and replication of the best-performing intervention is tested in the BEA. Specifically, at least two scores would be available for the best-performing intervention in addition to two scores from the comparison condition (i.e., an ineffective intervention or baseline), and the two within-condition averages could be compared to reduce SEM. The SEM for averages of two probes may still be large even after methods to select better performing probes are applied (e.g., 7-8 CWPM, Christ & Ardoin, 2009); for this reason, more than one replication may be needed to adequately demonstrate experimental control beyond expected, chance variation given the magnitude of SEMs on commonly-used CBM probe sets.

The use of CBM probes with approximately equal difficulty is essential for valid BEA inferences, increasing the likelihood that observed differences in performance are due to differential intervention effectiveness rather than differential probe difficulty. Fortunately, several methods to select probes based on difficulty have been developed. Of the available

methods, selecting a subset of probes based on mean performance and Euclidian distance (ED; a measure of dissimilarity for students' scores on two probes) has resulted in improvements in reliability for absolute decisions and SEM (Christ & Ardoin, 2009; Poncy et al., 2005). The utility of these methods to field test CBM probes prior to use in BEA has not been examined.

Current Study

The primary goal of this project is to determine the impact of controlling probe difficulty on reading BEA results. A group of first grade-level probes was field tested to identify two subsets of probes based on variation in difficulty—the least variable and most variable probes as determined by ED (see Christ & Ardoin, 2009). Following the field testing, BEA-like tasks (i.e., differentiating the effectiveness of two interventions in a BCBC design) were performed across the two probe sets. We hypothesized that best-performing interventions in BEA would be more likely to be replicated for individual students (a) when using the less variable CBM probe set and (b) when there are larger differences between students' performance on the interventions. We also hypothesized that (c) the impact of using less variable CBM probes would vary depending on the magnitude of differences in performance between interventions (i.e., that differences in the likelihood of replication between probe sets may be greater when there are smaller differences in performance between interventions). By investigating the consistency of results within and across probe sets, this study addressed the importance of using difficulty-controlled CBM probes. Specifically, if BEA results were inconsistent across probe sets and more consistent during the intervention replications on the difficulty-controlled passages, the data would support the importance of using difficulty-controlled probes in reading BEA.

Method

Participants

Students in 1st (*n* = 21) and 2nd (*n* = 20) grade at one elementary school in the Southern U.S. participated in the initial, field-testing phase of the study. Because the goal was to gather data from students across a wide range of skill levels, no specific inclusion criteria were specified at this phase of the study. Of the 41 participants, 17 (41%) were male. The most frequently reported student ethnic identifications, based on school records, were African American (56%) and Caucasian (39%), which is consistent with overall school demographics. Approximately 89% of students in the school qualified for free or reduced school lunches at the time of the study.

For the second phase of the study (i.e., the BEA-like tasks), participants were 40 students in 2nd (*n* = 14), 3rd (*n* = 7), 4th (*n* = 12), or 5th (*n* = 7) grade at a different elementary school in the same school district used in the first phase of this study. These students were selected based on teacher and principal nomination as students with reading difficulties. One participant was classified as a student with Specific Learning Disability, and the other participants had no current special education classification. Of the 40 participants, 78% were male, and 95% were identified as Caucasian in school records. Approximately 72% of students qualified for free or reduced lunches at the time of the study.

Measures

Sixteen, commercially available 1st grade reading probes were used in the study (Howe & Shinn, 2002). Of the 21 probes available at this grade level, 16 were selected to maximize variability in difficulty based on the values for mean performance reported in the technical manual (Howe & Shinn, 2002). For example, when considering probes for inclusion in the study,

only one probe was selected if two or more probes had very similar reported means in the technical report. All passages had high reported alternate-form correlations (>.80) and readability statistics calculated on the passages had high reported zero-order correlations (Howe & Shinn, 2002). Standard Aimsweb scoring procedures were used (Shinn & Shinn, 2002), and words correct per minute (WCPM) were scored on all passages.

Procedure

Phase 1.

In this phase, 16 reading probes were individually administered to students in counterbalanced sequence. These administrations occurred in short sessions over two to three consecutive days at the same time of day for individual students. Students received small incentives (e.g., colorful pencils, stickers) for participation. ED was calculated as a measure of dissimilarity between scores on each pair of probes (i.e., the square root of the sum, across students, of the squared deviations between each student's scores on the two probes). The four probes with the lowest mean ED across all pairwise probe comparisons were selected as the low variability probe set, and the four probes with the highest mean ED were selected as the high variability probe set.

The reliability of the two probe sets was compared based on calculation of generalizability and dependability coefficients (Shavelson & Webb, 1991), which assess reliability for relative and absolute decisions, respectively. In addition, the probe sets were compared based on calculation of SEM. In this study, overall variance in students' scores across probes within probe sets was considered to be a function of individual differences in student reading skill (*person*), variability in overall probe difficulty (*probe*), and differences in performance on specific probes by individual students (*residual*). These three variance components (*person, probe*, *residual*)

were estimated as random effects in linear mixed models using restricted maximum likelihood estimation (REML) with the 'lmer' function of the 'lme4' package (Bates, Maechler, & Bolker, 2011) in R (R Development Core Team, 2010). The REML estimator was used because simulation studies have supported REML as less biased than the analysis of variance approach (Marcoulides, 1990). To determine the generalizability coefficient, the variance component for *person* was divided by the sum of the *person* and *residual* variance components. To simulate what reliability would be for averages of more than one probe, the *residual* component was divided by the number of probes to be considered (i.e., two in this study) in the prior formula. Variance due to variations in probe difficulty (*probe*) was not included in the calculation of the generalizability coefficient because relative decisions (e.g., rank-ordering students) are not impacted by this source of variance. To determine the dependability coefficient, *person* was divided by the sum of *person*, *probe*, and *residual* variance. To simulate what reliability would be for averages of more than one probe, both the *probe* and *residual* components were divided by the number of probes to be used (i.e., two) and were substituted in the prior calculation. Variance related to differences in probe difficulty (*probe*) was included in this calculation because it would impact the absolute value of CBM scores and comparisons of specific scores to each other. The SEM was calculated by taking the square root of the appropriate estimate of error variance (*residual* or [*probe* + *residual*]), depending on whether relative or absolute comparisons were to be made.

Phase 2.

In this phase, students completed BEA-like tasks. Specifically, students completed two trials of two reading interventions on each probe set. In repeated reading (RR; Roshotte & Torgesen, 1985), students read the probe three times in succession, with CWPM scored on the third reading. In listening passage preview (Daly & Martens, 1994), the examiner read the probe aloud while the student followed along, and then CWPM was scored on the student's first reading. Each intervention was conducted twice on each probe set in alternating sequence (i.e.,

BCBC or CBCB). The order of the interventions was counterbalanced across students and probe sets. Students' CWPM on each intervention trial within probe set was compared to determine if the best-performing intervention was replicated (i.e., if [B1 > C1 and B2 > C2] or [B1 < C1 and B2 < C2], then results were considered to be replicated). In addition, average CWPM on RR ([B1 + B2]/2) and LPP ([C1 + C2]/2) within probe set was calculated, with the difference between these averages recorded for subsequent analyses.

To determine if the likelihood of replicating BEA results within probe set differed across probe sets and as a function of each student's difference in performance between RR and LPP trials, a generalized linear mixed model was fit with (a) probe set (0 = Low Variability, 1 = High Variability), (b) the within-probe set difference between average performance on LPP and RR, and (c) an interaction term (probe set X difference) as predictors of the likelihood of replicating results within probe set (0 = not replicated, 1 = replicated). A student-level random intercept was included to account for repeated measurements (i.e., that replication was assessed on two probe sets for each student). The binary distribution of the dependent variable was handled via a binomial logit link function in the model. These analyses were conducted the using 'lmer' function of the 'lme4' package (Bates et al., 2011) in R (R Development Core Team, 2010).

Inter-scorer Agreement and Procedural Integrity

All examiners were school psychology doctoral students who had completed formal coursework and training in CBM, the interventions used in this study (LPP and RR), and academic BEA prior to the study. In addition, all examiners had previously conducted multiple academic BEAs as part of intervention cases. During Phase 1, student performance across the 16 probes was audio recorded for 20% of all students. Student performance on each probe was scored, based on the recording, by a second examiner. Average agreement between the two

examiners across the 16 probes was high (97%). During Phase 2, agreement was assessed on 15% of all probes administered and inter-scorer agreement continued to be high (99%). Regarding procedural integrity, 16 intervention trials (RR or LPP) were fully recorded and reviewed. On 100% of the recorded trials, experimenters presented the probes in the correct sequence, had the student read the probes twice prior to the scored reading during RR, and read the probe to the student prior to the scored reading during LPP.

Results

Probe Set Selection

Descriptive statistics and the average ED for each probe is presented in Table 1. In general, scores on all probes exhibited some degree of positive skew and negative kurtosis. The four probes with the smallest average ED values were 1P14, 1P02, 1P08, and 1P05; consequently, these probes served as the Low Variability set. The four probes with the largest average ED values were 1P22, 1P04, 1P07, and 1P15, and these probes were identified as the High Variability set. Average scores on the Low Variability set ranged from 48.46 to 49.24 CWPM; in contrast, average scores on the High Variability set ranged from 44.66 to 56.29 CWPM.

The reliability and SEM of the probe sets were also compared, with variance components used in these calculations, indexes of generalizability (ρ^2) and dependability (ϕ), and SEM presented in Table 2. Of the presented indicators of reliability, the index of dependability and magnitude of SEM are of most importance because specific CBM scores are compared to other CBM scores (i.e., absolute decisions) in academic BEA.

Reliability for absolute decisions based on single probes was excellent for both probe sets (High Variability = .92, Low Variability = .98); however, there were substantial differences in SEM across the probe sets. For absolute decisions based on comparisons of single probes, the SEM was 5.76 on the Low Variability set compared to 12.17 on the High Variability set. In general, reliability and SEM improved for decisions comparing averages of two probes versus single probes.

BEA Replication

Overall, the probability of replicating the best-performing intervention within probe set was at near chance levels regardless of probe set variability; results were replicated for only 17 out of 40 students (43%) on both probe sets. The results of the hypothesized generalized linear mixed model, including the interaction term, are presented as Model 1 in Table 3. Results indicated that there were no differences across probe sets in the likelihood of replication (p = .84), with the best-performing intervention more likely to be replicated as the magnitude of differences in performance between RR and LPP increased ($p < .05$). There was a trend toward greater impact of differences between RR and LPP on the High Variability probes; however, this interaction term was not statistically significant (p = .09). Results excluding the interaction term (Model 2) were largely consistent with Model 1, again emphasizing the importance of the magnitude of differences between LPP and RR on the likelihood of replication ($p < .001$).

To illustrate effect size, the predicted odds and odds ratios at various values of differences between RR and LPP were determined. Based on Model 2 results, the predicted odds of replication were equal to $e^{(-1.78 + \text{difference}*.15)}$. Using this formula, the predicted odds of replication would be .36 for students with a 5-point difference between interventions and .76

for students with a 10-point difference, indicating that the odds of replication would be 2.11 as large for a 10-point as compared to a 5-point difference in performance across interventions.

Use of Average Scores in BEA

Because the probability of replication was low, even on the Low Variability probe set, we conducted exploratory analyses to determine the probability of replication when comparing averages of two probes. In these analyses, replication was assessed by comparing, across probe sets, the best performing intervention based on the within-probe set averages of RR ([RR1+RR2]/2) and LPP ([LPP1+LPP2]/2). For example, if a student's average RR score was greater than the average LPP score on both the High Variability and Low Variability probe sets, we judged results to be replicated in the BEA. Using this criterion, BEA results were replicated for 24 out of 40 students (60%), in comparison to the 43% replication rate when basing decisions on comparisons of individual probes. Differences in the frequencies of replication vs. non-replication by criterion (i.e., comparison of averages of two probes vs. single probes to assess replication) were statistically significant, as indicated by a chi-square test with Yates correction, $\chi^2(1) = 4.33$, $p < .05$. This improvement in the probability of replication is consistent with the reduction in SEM for absolute decisions when basing decisions on single vs. averages of two probes (see Table 2), and it is possible that the probability of replication could be even greater if the comparison of averages was not made across probe sets with variable SEMs, as is the case in this exploratory analysis. For absolute decisions based on averages of two probes, SEM was 4.07 in the Low Variability set, in contrast to the SEM of 8.61 in the High Variability set.

Discussion

In general, results were consistent with prior research indicating that there are substantial differences in reading CBM passage difficulty even on commercially available probes (e.g., Poncy et al., 2005). Although reliability for absolute decisions based on scores from single probes was excellent on both the least and most variable set of probes in this study, SEM varied considerably by probe set (~5 to 12 CWPM). These differences in SEM across probe sets had minimal impact on the probability of replicating students' BEA results when individual probes were considered. Rather, the primary determinant of replication in BEA was the magnitude of difference, on average, between RR and LPP trials for individual students. In other words, better differentiation in performance between the interventions for individual students improved the likelihood that consistent results would be obtained during the replication phase of the BEAs.

Because probe set quality, as determined by variability in passage difficulty, had no impact on the likelihood of replication, exploratory analyses were conducted to determine if replication would be more likely when averages of two probes per intervention condition were compared to determine the most effective intervention. Based on the reliability analyses, SEM was reduced for decisions based on averages of two vs. single probes, particularly on the more variable probe set. In addition, students likely exhibited some variability in their response to individual interventions; consequently, average scores across trials should improve estimation of students' true response to a particular intervention. When averages of two trials per intervention were compared, replication occurred in 60% of students' BEAs, in contrast to 43% when comparing individual CBM scores. Although these results are exploratory, they suggest

that comparison of average scores across intervention trials may be useful as a method to evaluate BEA results.

Implications for Theory and Practice

Currently, there are no consistent rules for determining the best-performing interventions in BEA (Burns & Wagner, 2008); however, there is a general trend over time toward greater numbers of trials per intervention in academic BEA studies. Although single intervention trials were administered with the best-performing intervention compared to baseline in a mini-withdrawal in one of the earlier reading BEA studies (Daly, Martens, Dool, & Hintze, 1998), the inclusion of 2 or more trials per intervention is common in more recent studies (e.g., Jones et al., 2009; McComas et al., 2009). This inclusion of more trials per intervention likely reflects the difficulty of differentiating interventions based on CBM data. Clear differentiation of performance across interventions is the primary criterion for evaluating multi-element or alternating treatment designs by visual analysis (Cooper, Heron, & Heward, 2007); however, it is possible that raters could disagree on the amount of differentiation necessary to demonstrate differences between interventions, particularly given variability in performance introduced by SEM of CBM and inconsistencies in students' responses to particular interventions.

To reduce this concern, comparison of averages across intervention trials to determine intervention effectiveness may be useful in developing a consistent, empirically-supported criterion for evaluating the best-performing intervention in academic BEA. One possibility could be to compare the magnitude of differences between intervention averages relative to published SEM, based on the number of averaged probes, for specific probe sets. As in

inferential statistics, the greater the magnitude of the observed differences relative to the expected differences based on SEM, the more confidence one would have that performance on the interventions has been adequately differentiated. Another option could be formal calculation of statistics based on the actual data in the BEA; however, given the limited number of observations per intervention condition, it may be difficult to obtain an adequate representation of within-intervention variability. In addition, as intervention trials are added to the BEA, autocorrelation, which is the degree to which subsequent scores can be predicted by prior scores, is likely to increase. For example, regardless of the specific interventions used, one would typically expect gradual improvement in skills over time, and this gradual, increasing trend would manifest as positive autocorrelation (Matyas & Greenwood, 1997). Because statistics such as *t*-tests and analysis of variance have assumptions of data independence, it is possible that autocorrelation could complicate the application of formal statistical calculations to BEA data.

Limitations

This study has several limitations that should be considered. Although students were referred by teachers and school principals as demonstrating limited reading skills, reading level was not confirmed prior to inclusion in the study. Given that the primary focus of the study was the ability to replicate results in academic BEA rather than specifically investigating the best-performing interventions for struggling readers, we believe that this limitation minimally impacted results, although could possibly limit the generalizability of findings.

In addition, the BEA-like task in this study included fewer conditions than a typical BEA. Generally, academic BEA includes baseline conditions in addition to several interventions of

varying complexity. Despite the simplicity of the design in this study, determining if there are differences in performance between two different conditions and if the differences replicate is a basic decisional task embedded in more complex BEA designs.

Summary and Future Directions

Overall, results indicated that the ability to replicate the best-performing intervention in academic BEA was minimally impacted by reducing variation in probe difficulty, at least based on subsets of the most and least variable probes on a commercially-available probe set. Of concern, replication remained at near chance levels even on the low variability probes that had near-optimal reliability levels and a SEM below most currently available reading CBM probe sets. Given these results, it appears that variation in student performance across different trials of specific interventions rather than psychometric properties of the probes themselves is the main source of random variance in academic BEA; however, basing decisions on average performance across multiple trials of the same intervention can improve decisions by reducing measurement-related SEM, as well as providing better estimates of students' true response to the intervention.

Based on these results, several areas warrant further investigation. First, the generalizability of these findings to BEAs with more than two trials per intervention needs to be assessed. In such a study, improvement in the likelihood of replication as additional intervention trials are added can be examined relative to costs in efficiency of administration. Second, the performance of criteria regarding the magnitude of mean differences between interventions, relative to expected SEM based on prior reliability analyses, can be investigated. It is possible that these criteria, which would not require complex statistical calculations beyond

the initial reliability studies, could improve decision-making without greatly adding to the complexity of the BEA process. Given the limited use of and perceived need for statistical analysis in single-case research and practice (Perone, 1999), this approach, provided it proves to be technically sound, may be more acceptable and accessible to researchers and practitioners than formal statistical analysis. Last, because some features of time-series data (e.g., autocorrelation) complicate most commonly-used statistics, the performance of these tests in the context of BEA needs to be examined in statistical simulations. It is possible that widely-used statistics such as *t*-tests may be minimally biased when analyzing BEA data, and the ability to use simple statistical analyses to support decisions could facilitate more widespread use in BEA studies and practice. Through these efforts, we hope that empirically-supported methods to guide BEA decisions are developed for use in research and practice.

References

Bates, D., Maechler, M., & Bolker, B. (2011). *lme4: Linear mixed-effects models using S4 classes*. R package (Version 0.999375-39). Retrieved from http://CRAN.R-project.org/package=lme4

Burns, M. K., Ganuza, Z. M., & London, R. M. (2009). Brief experimental analysis of written letter formation: Single-case demonstration. *Journal of Behavioral Education, 18*, 20-34. doi: 10.1007/s10864-008-9076-z

Burns, M. K., & Wagner, D. (2008). Determining an effective intervention within a brief experimental analysis for reading: A meta-analytic review. *School Psychology Review, 37*, 126-136.

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

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Upper Saddle River, NJ: Pearson.

Daly, E. J., III, Bonfiglio, C. M., Mattson, T., Persampieri, M., & Foreman-Ya, K. (2006). Refining the experimental analysis of academic skills deficits: Part II. use of brief experimental analysis to evaluate reading fluency treatments. *Journal of Applied Behavior Analysis, 39*, 323-331.

Daly, E. J., III, & Martens, B. K. (1994). A comparison of three interventions for increasing oral reading performance: Application of the instructional hierarchy. *Journal of Applied Behavior Analysis, 27*, 459-469. doi: 10.1901/jaba.1994.27-459

Daly, E. J., III, Martens, B. K., Dool, E. J., & Hintze, J. M. (1998). Using brief functional analysis to select interventions for oral reading. *Journal of Behavioral Education, 8*, 203-218. doi: 10.1023/a:1022835607985

Daly, E. J., III, Persampieri, M., McCurdy, M., & Gortmaker, V. (2005). Generating reading interventions through experimental analysis of academic skills: Demonstration and empirical evaluation. *School Psychology Review, 34*, 395-414.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53*, 199-208.

Good, R. H., & Kaminski, R. A. (2002). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement.

Hintze, J. M., Owen, S. V., Shapiro, E. S., & Daly, E. J., III. (2000). Generalizability of oral reading fluency measures: Application of G theory to curriculum-based measurement. *School Psychology Quarterly, 15*, 52-68. doi: 10.1037/h0088778

Howe, K. B., & Shinn, M. M. (2002). *Standard reading assessment passages (RAPs) for use in general outcome measurement: A manual describing development and technical features.* Retrieved from http://www.aimsweb.com

Jones, K. M., Wickstrom, K. F., & Daly, E. J., III. (2008). Best practices in the brief assessment of reading concerns. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (5th ed., Vol. 2, pp. 489-501). Bethesda, MD: National Association of School Psychologists.

Jones, K. M., Wickstrom, K. F., Noltemeyer, A. L., Brown, S. M., Schuka, J. R., & Therrien, W. J. (2009). An experimental analysis of reading fluency. *Journal of Behavioral Education, 18*, 35-55. doi: 10.1007/s10864-009-9082-9

Kratochwill, T. R., & Stoiber, K. C. (2002). Evidence-based interventions in school psychology: Conceptual foundations of the Procedural and Coding Manual of Division 16 and the Society for the Study of School Psychology Task Force. *School Psychology Quarterly, 17*, 341-389. doi: 10.1521/scpq.17.4.341.20872

Marcoulides, G. A. (1990). An alternative method for estimating variance components in generalizability theory. *Psychological Reports, 66*, 379-386. doi: 10.2466/pr0.66.2.379-386

Matyas, T. A., & Greenwood, K. M. (1997). Serial dependency in single-case time series. In R. D. Franklin, D. B. Allison & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 215-243). Mahwah, NJ: Erlbaum.

McComas, J. J., Wagner, D., Chaffin, M. C., Holton, E., McDonnell, M., & Monn, E. (2009). Prescriptive analysis: Further individualization of hypothesis testing in brief experimental analysis of reading fluency. *Journal of Behavioral Education, 18*, 56-70. doi: 10.1007/s10864-009-9080-y

Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *The Behavior Analyst, 22*, 109-116.

Petursdottir, A., McMaster, K., McComas, J. J., Bradfield, T., Braganza, V., Koch-McDonald, J., . . . Scharf, H. (2009). Brief experimental analysis of early reading interventions. *Journal of School Psychology, 47*, 215-243. doi: 10.1016/j.jsp.2009.02.003

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

R Development Core Team. (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Roshotte, C. A., & Torgesen, J. K. (1985). Repeated reading and reading fluency in learning disabled children. *Reading Research Quarterly, 20*, 180-188. doi: 10.1598/rrq.20.2.4

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*: Sage.

Shinn, M. M., & Shinn, M. R. (2002). *Administration and scoring of reading curriculum-based measurement (R-CBM) for use in general outcome measurement.* Retrieved from http://www.aimsweb.com

VanDerHeyden, A. M., & Burns, M. K. (2009). Performance indicators in math: Implications for brief experimental analysis of academic performance. *Journal of Behavioral Education, 18*, 71-91. doi: 10.1007/s10864-009-9081-x

Table 1

Average Euclidean distance and descriptive statistics by probe

Probe Number	ED	*M*	*SD*	Skew[a]	Kurtosis[b]
1P01	73.84	49.54	38.95	.34	-1.22
1P02[LV]	65.67	48.46	40.33	.58	-.96
1P04[HV]	93.46	58.73	44.71	.26	-1.45
1P05[LV]	68.92	49.24	41.39	.55	-1.15
1P07[HV]	85.74	44.66	36.16	.38	-1.37
1P08[LV]	67.79	49.24	37.67	.34	-1.21
1P09	78.53	47.29	42.98	.51	-1.19
1P10	70.81	50.90	38.17	.30	-1.28
1P11	70.34	51.10	42.81	.44	-1.20
1P13	78.17	44.80	38.75	.49	-1.12
1P14[LV]	64.15	49.20	40.72	.47	-1.22
1P15[HV]	82.85	51.90	45.69	.37	-1.27
1P16	75.87	53.54	41.39	.44	-1.12
1P17	73.70	49.80	41.87	.72	-.65
1P19	75.02	51.66	37.71	.35	-1.17
1P22[HV]	117.66	56.29	44.63	.41	-1.21

Note. HV = High Variability probe set, LV = Low Variability probe set, [a]*SE* = .37, [b]*SE* = .72

Table 2

Reliability by probe set and number of probes

	Low Variability		High Variability	
Source of Variance	1 Probe	2 Probes	1 Probe	2 Probes
Person	1570.43	1570.43	1733.62	1733.62
Probe	--	--	35.39	17.70
Residual	33.13	16.57	112.80	56.40
Reliability				
ρ^2	.98	.99	.94	.97
Φ	.98	.99	.92	.96
SEM				
Δ	5.76	4.07	10.62	7.51
Δ	5.76	4.07	12.17	8.61

Note. ρ^2 = index of generalizability (relative decisions), ϕ = index of dependability (absolute decisions), Δ = SEM for relative decisions, δ = SEM for absolute decisions.

Table 3

Results of generalized linear mixed models

	Model 1			Model 2		
Fixed Effects	Estimate	*SE*	*p*	Estimate	*SE*	*p*
Intercept	-1.21	.58	.04	-1.78	.54	.00
Probe Set	-1.67	1.08	.12	-.10	.51	.84
Difference	.09	.05	.05	.15	.04	.00
Probe Set*Difference	.15	.09	.09	--	--	--

Note. Probe Set was coded as 0 = Low Variability probe set, 1 = High Variability probe set.