School Psychology Quarterly
2012, Vol. 27, No. 4, 223–235

1045-3830/12/$12.00 DOI: 10.1037/spq0000001

The Impact of Probe Variability 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

The purpose of this study was to examine the impact of probe variability on the ability to replicate results in brief experimental analysis (BEA) of reading. In the first phase of the study, 41 first- and second- grade students completed 16 oral reading fluency probes. Calculations of probe difficulty were used to identify Low and High Variability probe sets. In the second phase of the study, the performance of 40 second- through fifth-grade students during two reading interventions was compared. The best-performing intervention for each student in the initial trial was replicated during a second trial for only 43% of students regardless of probe variability. The best-performing intervention was replicated for 60% of students when average performance across two trials was compared. Rules for determining the best-performing intervention in academic BEA should consider the standard error of measurement (SEM) for the probe set to be used, the reliability for absolute decisions using the probe set, and the number of replications relative to SEM needed to adequately demonstrate experimental control.

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

Educational professionals increasingly are encouraged to implement evidence-based instructional practices (Burns & Ysseldyke, 2009), 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 individual students. Logically, the likelihood of intervention success may be improved by selecting an intervention with the largest reported effect size, yet this strategy may not ensure that the best and most efficient intervention will be selected for a particular student. These concerns regarding intervention effectiveness can be ameliorated 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 multielement 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 miniwithdrawal (Jones, Wickstrom, & Daly, 2008). For example, in Jones, Wickstrom, and Daly (2008) correct words per

Sterett H. Mercer, Department of Educational and Counselling Psychology and Special Education, University of British Columbia, Vancouver, British Columbia, Canada; 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. Portions of this article were presented at the 2012 annual meeting of the National Association of School Psychologists.

Correspondence concerning this article should be addressed to Sterett H. Mercer, 2125 Main Mall, Vancouver, BC V6T 1Z4 Canada. E-mail: sterett.mercer@ubc.ca

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 a miniwithdrawal. 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). The effectiveness of BEA has been supported in the literature, with several studies reporting that initial BEA results for best-performing interventions were confirmed in extended analyses and that effect sizes for BEA-identified interventions were larger relative to other intervention conditions for individual students (for a review, see Burns & Wagner, 2008).

Reliability of BEA Decisions

Although BEA methods have been adapted and applied to diverse academic skills such as reading fluency (e.g., Daly, Bonfiglio, Mattson, Persampieri, & Foreman-Ya, 2006), letter sound fluency (e.g., Petursdottir et al., 2009), math computation (e.g., VanDerHeyden & Burns, 2009), and basic writing skills (e.g., Burns, Ganuza, & London, 2009), the reliability of BEA decisions needs to be addressed prior to more widespread use of BEA methods. There have been few efforts to formally define and consider reliability in the context of BEA designs, yet replication of results is fundamental to BEA designs and necessary for demonstration of experimental control (Martens, Eckert, Bradley, & Ardoin, 1999). Consequently, replication of results would appear to serve as a reasonable indicator of reliability in BEA. Confidence in BEA results increases proportionally to the number of replications of intervention effects in BEAs (Martens et al., 1999); however, there are no accepted guidelines regarding the number of replications that are needed to demonstrate experimental control (Burns & Wagner, 2008). At minimum, replication through the use of one miniwithdrawal is necessary to demonstrate experimental control, and some academic BEA designs add additional criteria, such as a 20–30% increase compared with baseline during both the initial trial and miniwithdrawal to ensure that intervention effects are substantial (Jones et al., 2009; Noell, Freeland, Witt, & Gansle, 2001).

Several features of academic BEA indicate that replication in one miniwithdrawal may be insufficient to support the reliability of the BEA decision. Namely, most academic BEA methods incorporate curriculum-based measures (CBM) of academic skills, and decisions using CBM data can have limited reliability under certain conditions. Early research on oral reading CBM (Derr & Shapiro, 1989; Derr-Minneci & Shapiro, 1992) demonstrated that students' scores are sensitive to differences in examiner (i.e., teacher vs. school psychologist), testing location (i.e., reading group, teacher's desk, or office), and task demand (i.e., timed vs. untimed); consequently, it is important to keep testing conditions consistent in BEA to maximize reliability.

In addition to testing conditions, reliability of BEA decisions may be influenced by the reliability of the CBMs used in the BEA. 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 specific scores on prior occasions or cut-scores) as compared with relative decisions (rank-ordering scores without considering magnitude of differences). In academic BEA the primary judgments are determining (a) if a student's performance following one or more intervention trials is better than performance during baseline or following other interventions and (b) the magnitude of differences between conditions; consequently, these judgments incorporate absolute decisions. Ideally, guidelines for the number of replications needed in BEA to demonstrate ex-

perimental control should consider the reliability of absolute decisions when based on CBM data because measurement-related unreliability could generate the appearance of significant intervention effects when no true intervention effects are present, particularly when few probes have been administered in a limited number of intervention trials.

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, Skinner, and Axtell (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). Large magnitude SEM is particularly concerning given that SEM can inflate the standard error of progress monitoring slopes (Christ, 2006), complicating efforts to accurately determine growth rates and students' responsiveness to intervention.

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 probe set average. In addition, selecting optimal probes based on euclidean Distance (ED; a measure of dissimilarity for students' scores on two probes) has resulted in substantial improvements to reliability and SEM (Christ & Ardoin, 2009). Similar procedures have been used to select optimal probes for individuals in academic BEA studies—in Daly, Persampieri, McCurdy, and Gortmaker (2005), participants read 26 passages prior to the BEAs to identify participant-specific subsets of passages of approximately equal difficulty. However, conducting field testing procedures for each student prior to BEAs reduces the efficiency of the process.

Impact of SEM on BEA Decisions

Determining the magnitude of performance differences between intervention conditions based on a limited number of observations is central to BEA methodology; consequently, the reliability of CBM for absolute decisions has implications for the reliability of BEA decisions. The magnitude of SEM for absolute decisions is particularly relevant to BEA decisions given that SEM characterizes the expected variation around CBM scores—if the difference in performance between two conditions is not well beyond expected error variation, additional BEA trials are needed to reduce SEM and increase confidence that experimental control has been demonstrated. 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 (i.e., 2*SEM) without any true change in reading skill. This range of expected variation is reduced when considering averages of multiple probes, and averages of two or more probes are available for comparison so long as 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., a less effective intervention or baseline), and the two within-condition averages could be compared with 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 SEMs of 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. Some BEA studies have relied on readability estimates as evidence of comparable difficulty (Daly, Martens, Dool, & Hintze, 1998; Jones & Wickstrom, 2002; Noell et al., 2001), but readability estimates are poor

predictors of actual student performance on reading CBM (Ardoin, Williams, Christ, Klubnik, & Wellborn, 2010; Betts, Pickart, & Heistad, 2009). Student-specific screening of all passages considered for use in the BEA (e.g., Daly et al., 2005) provides stronger evidence of equivalence but reduces the efficiency of the BEA process. Consequently, additional group-level field testing of existing probe sets may be a more efficient method to establish comparability of probes to be used in BEA. Of the available field testing methods, selecting a subset of probes based on ED has best reduced SEM and maximally improved reliability (Christ & Ardoin, 2009); however, the utility of this method 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 potential impact of controlling probe difficulty on decision making in BEA. 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. Following the field testing, intervention trials (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 for individual students would be more likely to be replicated (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). Although the BCBC design used in this study included fewer components and interventions than would typically be included in a BEA, this design facilitated examination of a critical component of BEAs—the ability to replicate differences in performance between two experimental conditions.

Method

Participants

Students in first ($n = 21$) and second ($n = 20$) grade at one elementary school in the Southern U.S. participated in the field-testing phase of the study. School demographic records indicated that most students in the school were African American (61%) or Caucasian (36%), with the majority of students (71%) qualifying for free or reduced school lunches. Because the goal was to gather data from students across a wide range of skill levels, no specific inclusion criteria were specified other than enrollment in the target grades 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. The sample size of 41 was greater than the often recommended sample size of 30 for reasonably accurate estimation of variance components (Snijders & Bosker, 1999, p. 154).

For the second phase of the study (i.e., the intervention trials), participants were 40 students in second ($n = 14$), third ($n = 7$), fourth ($n = 12$), or fifth ($n = 7$) grade at a different elementary school in the same school district used in the first phase of this study. School demographic records indicated that most students in the school were Caucasian (94%), with the majority of students (51%) qualifying for free or reduced school lunches. These students were selected based on teacher and principal nomination as students with estimated reading skills at approximately the first grade level, but student reading levels were not confirmed prior to participation in this phase of the study. 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. With a sample size of 40, power was .91 to detect an odds ratio of 3:1 (probability of replication on less variable probes = .75, probability on highly variable probes = .50) with a within-subject correlation of .20 in a generalized linear mixed model (Dang, Mazumdar, & Houck, 2008).

Measures

Sixteen commercially available first grade reading probes were used in the study (Howe & Shinn, 2002). Of the 21 probes available at this grade level, 16 were selected to maintain initial 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 nearly identical reported means in the technical report. This approach for initial probe selection was used because we believed it would be more difficult to identify a high variability than a low variability subset of probes in a previously field-tested probe set. 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

Both phases of the project were approved by a university Institutional Review Board, and active written parental consent and verbal assent of students were obtained prior to participation.

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. Statistical analyses, described in a subsequent section, were used to identify Low Variability (LV) and High Variability (HV) probe sets.

Phase 2. In this phase, students completed two trials of two reading interventions on each probe set. Intervention trials were conducted in 20–30 minute sessions on different days (e.g., Day 1 = HV trials, Day 2 = LV trials), with the ordering of the HV and LV trials counterbalanced. 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 (LPP; 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. LPP and RR were selected as interventions because they are frequently included in reading BEA (Burns & Wagner, 2008). 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.

Interscorer 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 average interscorer 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.

Data Analysis

Phase 1. To identify the LV and HV probe sets, 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 LV probe set, and the four probes with the highest mean ED were selected as the HV 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, 2012). 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. To determine if the likelihood of replicating intervention 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 = *LV*, 1 = *HV*), (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, 2012).

Results

Probe Set Selection

Descriptive statistics and the average ED for each probe are 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 LV set. The four probes with the largest average ED values were 1P22, 1P04, 1P07, and 1P15, and these probes were identified as the HV set. Average scores on the LV set ranged from 48.46 to 49.24 CWPM; in contrast, average scores on the HV 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 with other CBM scores and the magnitude of differences is considered (i.e., absolute decisions) in academic BEA. Most assumptions of the linear

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. Sample size = 41 for all probes. HV = High variability probe set; LV = Low variability probe set.
[a] *SE* = .37. [b] *SE* = .72.

mixed models were met, including normality of residuals and the *person* and *probe* random effects. The assumption of equal variance for the residuals appeared questionable because residuals were less variable at lower CWPM. Violation of the homogeneity assumption is most problematic when the residuals vary as a function of predictors in the model (Raudenbush & Bryk, 2002), and there were no predictors in these analyses.

Reliability for absolute decisions based on single probes was excellent for both probe sets (HV = .92, LV = .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 LV set compared with 12.17 on the HV set. In general, reliability and SEM improved for decisions comparing averages of two probes versus single probes.

Intervention Replication

Mean CWPM by intervention, trial, and probe set is presented separately by student grade in Table 3. In addition, the proportion of

Table 2
Reliability by Probe Set and Number of Probes

Source of variance	Low variability		High variability	
	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
Mean Correct Words per Minute and Proportion of Replications by Intervention and Probe Set for Phase 2 Participants

Variable	Grade			
	2	3	4	5
LV: LPP Trial 1	76.79	95.43	125.20	161.29
LV: LPP Trial 2	71.43	100.36	124.80	157.43
LV: RR Trial 1	79.14	103.00	119.20	158.43
LV: RR Trial 2	76.50	104.07	125.00	165.14
HV: LPP Trial 1	72.21	103.29	146.60	155.86
HV: LPP Trial 2	70.43	104.00	135.60	150.71
HV: RR Trial 1	77.64	102.36	141.20	164.14
HV: RR Trial 2	77.00	107.57	158.20	163.29
Replication: HV	.43	.57	.40	.14
Replication: LV	.50	.57	.20	.14
Replication: Average scores	.64	.64	.60	.43
Sample size	14	14	5	7

Note. LV = Low variability; HV = High variability; LPP = Listening passage preview; RR = Repeated readings.

trials in which the best-performing intervention was replicated is presented by probe set and student grade in Table 3. 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. As displayed in Table 3, there was an apparent trend of lower replication likelihood on the LV and HV probe sets at higher student grades; however, these differences were not statistically significant, $F(4, 40) = 1.93$, $p = .12$. Results of the hypothesized generalized linear mixed model, including the interaction term, are presented as Model 1 in Table 4. Generalized linear mixed models do not have homogeneity or normality assumptions due to the use of a link function to transform responses on the dependent variable (Raudenbush & Bryk, 2002). 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 HV probes; however, this interaction term was only a statistical trend ($p = .09$). Because the coefficient for probe set was negative and the coefficients for differences and the probe set by differences interaction term were positive, the interaction term, if statistically significant, would suggest that the reduced likelihood of replication on the HV probe set is mitigated as the magnitude of differences between RR and LPP increases. Results excluding the interaction term (Model 2) were largely consistent with

Table 4
Results of Generalized Linear Mixed Models

Fixed effects	Model 1			Model 2		
	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.

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 with a 5-point difference in performance across interventions.

Use of Average Scores in BEA

Because the probability of replication was low, even on the LV 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 HV and LV probe sets, we judged results to be replicated in the BEA. Using this criterion, intervention 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, and replication rates by grade are presented in Table 3. Differences in the frequencies of replication versus nonreplication 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 versus 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 LV set, in contrast to the SEM of 8.61 in the HV set.

Discussion

The primary purpose of the study was to investigate the potential impact of probe variability on decision making in BEA by determining the likelihood of replicating the best-performing intervention in a BCBC design. Contrary to our primary hypothesis, there were no significant differences in replication rates between the LV and HV probe sets. In contrast, results were consistent with the hypothesis that the likelihood of replicating intervention results would be greater with larger differences in student performance between the two interventions, and there was a statistical trend toward fewer differences in replication rates between the LV and HV probe sets as the magnitude of differences in student performance between the interventions increased. Consequently, the primary determinant of the likelihood of replication was the magnitude of separation in performance across interventions for each student.

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 LV and HV sets 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' intervention results when individual probes were considered. Rather, the primary determinant of replication was the magnitude of difference, on average, between RR and LPP trials for individual students. In other words, better differentiation in performance, relative to the magnitude of SEM, between the interventions for individual students improved the likelihood that consistent results would be obtained during the replication phase and that preliminary evidence of experimental control would be demonstrated.

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 versus single probes, particularly on the HV 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' intervention trials, in contrast to 43% when comparing individual CBM scores. Averages of only two trials were investigated in this study due to the study design, but the reliability for estimating differences in performance between interventions will generally increase as additional trials per intervention are added. 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 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 with baseline in a mini-withdrawal in one of the earlier reading BEA studies (Daly et al., 1998), the inclusion of two 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 too few CBM probes. Clear differentiation of performance across interventions is the primary criterion for evaluating multielement 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. This concern is partially mitigated when criteria such as 20–30% differences in performance between conditions are required in the BEA design (e.g., Jones et al., 2009; Noell et al., 2001); although the 20–30% requirements were not based on reliability considerations, greater differentiation between conditions would reduce the likelihood that the differences occurred due to measurement error.

To reduce the concern that observed differences may be an artifact of measurement error, 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 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, may increase. For example, regardless of the specific interventions used, one might 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. Consequently, comparing the magnitude of differences between intervention conditions relative to estimates of SEM based on the probe set used and number of replications in the BEA may be a more viable option to support that experimental control has been adequately demonstrated.

Limitations

This study has several limitations that should be considered. Although students were referred by teachers and school principals as demonstrat-

ing reading skills at approximately a first grade level, reading skills were not confirmed prior to inclusion in the study. Based on the mean CWPM by grade presented in Table 3, many students, particularly those enrolled in higher grades, appeared to be instructional at levels beyond the first grade. Although there was a general trend of less replication at higher student grades, the differences between grades were not statistically significant, suggesting that student skill level did not substantially impact results. Verification of student skill levels would have ensured that the probes used in the study were sensitive to instructional manipulations, and future studies would benefit from more rigorous inclusion criteria.

In addition, the intervention trials 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. Future research should examine the generalizability of the current study's results to more complex designs.

Last, field testing in this study demonstrated that the SEM of commercially available probes can be improved; however, the use of probes that have undergone prior field testing during development may have reduced differences in reliability and SEM between the LV and HV probe sets in this study. Regardless, there was a low probability of replication of the best-performing interventions, even with SEM on the LV set below most available probe sets, stressing the importance of extended analysis in academic BEA.

Summary and Future Directions

Overall, results indicated that the ability to replicate the best-performing intervention 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 LV 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 may be 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

Ardoin, S. P., Williams, J. C., Christ, T. J., Klubnik, C., & Wellborn, C. (2010). Examining readability estimates' predictions of students' oral reading

rate: Spache, Lexile, and Forcast. *School Psychology Review, 39,* 277–285.

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

Betts, J., Pickart, M., & Heistad, D. (2009). An investigation of the psychometric evidence of CBM-R passage equivalence: Utility of readability statistics and equating for alternate forms. *Journal of School Psychology, 47,* 1–17. doi:10.1016/j.jsp.2008.09.001

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.

Burns, M. K., & Ysseldyke, J. E. (2009). Reported prevalence of evidence-based instructional practices in special education. *The Journal of Special Education, 43,* 3–11. doi:10.1177/0022466908315563

Christ, T. J. (2006). Short-term estimates of growth using curriculum-based measurement of oral reading fluency: Estimating standard error of the slope to construct confidence intervals. *School Psychology Review, 35,* 128–133.

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.

Dang, Q., Mazumdar, S., & Houck, P. R. (2008). Sample size and power calculations based on generalized linear mixed models with correlated binary outcomes. *Computer Methods and Programs in Biomedicine, 91,* 122–127. doi:10.1016/j.cmpb.2008.03.001

Derr, T. F., & Shapiro, E. S. (1989). A behavioral evaluation of Curriculum-Based Assessment of reading. *Journal of Psychoeducational Assessment, 7,* 148–160. doi:10.1177/073428298900700205

Derr-Minneci, T. F., & Shapiro, E. S. (1992). Validating curriculum-based measurement in reading from a behavioral perspective. *School Psychology Quarterly, 7,* 2–16. doi:10.1037/h0088244

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. (2002). Done in sixty seconds: Further analysis of the Brief Assessment Model for academic problems. *School Psychology Review, 31,* 554–568.

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

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

Martens, B. K., Eckert, T. L., Bradley, T. A., & Ardoin, S. P. (1999). Identifying effective treatments from a brief experimental analysis: Using single-case design elements to aid decision making. *School Psychology Quarterly, 14,* 163–181. doi:10.1037/h0089003

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

Noell, G. H., Freeland, J. T., Witt, J. C., & Gansle, K. A. (2001). Using brief assessments to identify effective interventions for individual students. *Journal of School Psychology, 39,* 335–355. doi:10.1016/s0022-4405(01)00072-3

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd ed.)*. Thousand Oaks, CA: Sage.

R Development Core Team. (2012). *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*. Thousand Oaks, CA: Sage.

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

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London, UK: Sage.

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

Received July 30, 2012
Accepted August 7, 2012 ■

UNITED STATES POSTAL SERVICE ® **Statement of Ownership, Management, and Circulation (All Periodicals Publications Except Requester Publications)**

1. Publication Title	2. Publication Number	3. Filing Date
School Psychology Quaererly	0 0 7 - 6 5 9	October 2012

4. Issue Frequency	5. Number of Issues Published Annually	6. Annual Subscription Price
Quarterly	4	Mbr $65 Indiv $120 Inst $454

7. Complete Mailing Address of Known Office of Publication (Not printer) (Street, city, county, state, and ZIP+4®): 750 First Street, N.E., Washington, D.C. 20002-4242 — Contact Person: Barbara Spruill, Telephone (Include area code): 202 136-5579

8. Complete Mailing Address of Headquarters or General Business Office of Publisher (Not printer): 750 First Street, N.E., Washington, D.C. 20002-4242

9. Full Names and Complete Mailing Addresses of Publisher, Editor, and Managing Editor (Do not leave blank)

Publisher (Name and complete mailing address): Educational Publishing Foundation/APA, 750 First Street, N.E. Washington, D.C. 20002-4242

Editor (Name and complete mailing address): Randy Kamphaus, Dean, College of Education, Georgia State University, 30 Pryon Street, Atlanta, GA 30302-3965

Managing Editor (Name and complete mailing address): Susan J.A. Harris, American Psychological Association, 750 First Street, N.E., Washington, D.C. 20002-4242

10. Owner (Do not leave blank. If the publication is owned by a corporation, give the name and address of the corporation immediately followed by the names and addresses of all stockholders owning or holding 1 percent or more of the total amount of stock. If not owned by a corporation, give the names and addresses of the individual owners. If owned by a partnership or other unincorporated firm, give its name and address as well as those of each individual owner. If the publication is published by a nonprofit organization, give its name and address.)

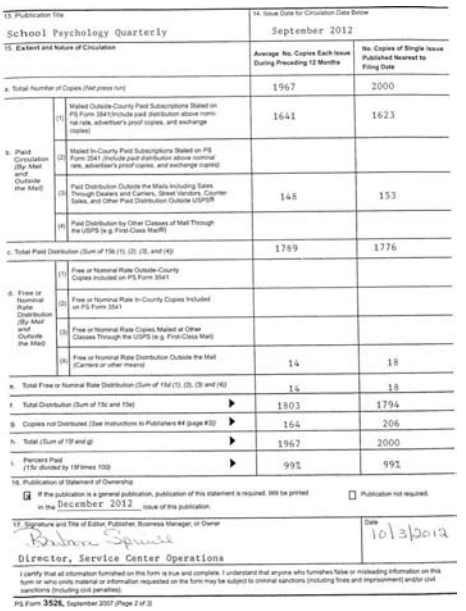
Full Name	Complete Mailing Address
American Psychological Association	750 First Street, N.E. Washington, D.C. 20002-4242

11. Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds, Mortgages, or Other Securities. If none, check box → ☒ None

Full Name	Complete Mailing Address
NONE	

12. Tax Status (For completion by nonprofit organizations authorized to mail at nonprofit rates) (Check one)
The purpose, function, and nonprofit status of this organization and the exempt status for federal income tax purposes:
☒ Has Not Changed During Preceding 12 Months
☐ Has Changed During Preceding 12 Months (Publisher must submit explanation of change with this statement)

PS Form **3526**, September 2007 (Page 1 of 3 (Instructions Page 3)) PSN 7530-01-000-9931 **PRIVACY NOTICE:** See our privacy policy on www.usps.com



13. Publication Title	14. Issue Date for Circulation Data Below
School Psychology Quarterly	September 2012

15. Extent and Nature of Circulation		Average No. Copies Each Issue During Preceding 12 Months	No. Copies of Single Issue Published Nearest to Filing Date
a. Total Number of Copies (Net press run)		1967	2000
b. Paid Circulation (By Mail and Outside the Mail)	(1) Mailed Outside-County Paid Subscriptions Stated on PS Form 3541(Include paid distribution above nominal rate, advertiser's proof copies, and exchange copies)	1641	1623
	(2) Mailed In-County Paid Subscriptions Stated on PS Form 3541 (Include paid distribution above nominal rate, advertiser's proof copies, and exchange copies)		
	(3) Paid Distribution Outside the Mails Including Sales Through Dealers and Carriers, Street Vendors, Counter Sales, and Other Paid Distribution Outside USPS®	148	153
	(4) Paid Distribution by Other Classes of Mail Through the USPS (e.g. First-Class Mail®)		
c. Total Paid Distribution (Sum of 15b (1), (2), (3), and (4))		1789	1776
d. Free or Nominal Rate Distribution (By Mail and Outside the Mail)	(1) Free or Nominal Rate Outside-County Copies included on PS Form 3541		
	(2) Free or Nominal Rate In-County Copies Included on PS Form 3541		
	(3) Free or Nominal Rate Copies Mailed at Other Classes Through the USPS (e.g. First-Class Mail)		
	(4) Free or Nominal Rate Distribution Outside the Mail (Carriers or other means)	14	18
e. Total Free or Nominal Rate Distribution (Sum of 15d (1), (2), (3) and (4))		14	18
f. Total Distribution (Sum of 15c and 15e)		1803	1794
g. Copies not Distributed (See Instructions to Publishers #4 (page #3))		164	206
h. Total (Sum of 15f and g)		1967	2000
i. Percent Paid (15c divided by 15f times 100)		99%	99%

16. Publication of Statement of Ownership
☒ If the publication is a general publication, publication of this statement is required. Will be printed in the December 2012 issue of this publication. ☐ Publication not required.

17. Signature and Title of Editor, Publisher, Business Manager, or Owner: Barbara Spruill, Director, Service Center Operations — Date: 10/3/2012

I certify that all information furnished on this form is true and complete. I understand that anyone who furnishes false or misleading information on this form or who omits material or information requested on the form may be subject to criminal sanctions (including fines and imprisonment) and/or civil sanctions (including civil penalties).

PS Form **3526**, September 2007 (Page 2 of 3)