academicresearchJournals

Vol. 4(2), pp. 29-46, February 2016

DOI: 10.14662/IJARER2016.001

ISSN: 2360-7866

http://www.academicresearchjournals.org/IJARER/Index.htm

International Journal of Academic Research in Education and Review

Full Length Research

An Examination of the Effects of Two Equating Methods on the Equivalence of the First Grade DIBELS Oral Reading Fluency Probes

[*1]Chung-Hau Fan, [2]Peter R. Denner, [3]Yi Ding and [4]Yu-Lin Chang

[*1]University of Iowa

[2]Idaho State University

[3]Fordham University

[4]National Taiwan Normal University

*Corresponding author's E-mail: fanchun@isu.edu

Accepted 24 February 2016

Research has revealed concerns regarding the limited psychometric evidence for using CBM oral reading fluency (CBM-R) for progress monitoring at the individual student level due to probe nonequivalence. Equating methods based on test theories have been shown to be useful to equate less comparable forms. This study investigated the effect of using mean equating and linear equating methods for managing score variability at the individual student level across DIBELS ORF (DORF) probes. A sample of 68 first grade students were administered the first grade DORF probes, and their words correctly read per min (WCPM) were calculated. The results indicated the comparability of the DORF outcomes was significantly improved with the two equating methods. Additionally, linear equating outperformed mean equating for managing score variability. Nevertheless, noticeable score variability was still observed at the individual student level. Directions for future study and implications for applying equating methods for making educational decisions are discussed.

Keywords: Oral reading fluency, equating, curriculum-based measurement, progress monitoring

Cite This Article As: Fan C-H, Denner PR, Ding Y, Chang Y-L (2016). An Examination of the Effects of Two Equating Methods on the Equivalence of the First Grade DIBELS Oral Reading Fluency Probes. Inter. J. Acad. Res. Educ. Rev. 4(2): 29-46

INTRODUCTION

Curriculum-based measurement (CBM) is an approach to measuring the academic growth of individual students frequently to help teachers in evaluating the effectiveness of their instruction (Deno, 1985). Historically, CBM data have been used to guide low-stakes decisions (e.g., responsiveness to classroom instruction, pre-referral intervention effectiveness; Shinn, 1998). The reauthorization of the Individuals with Disabilities Education Improvement Act (IDEIA, 2004) allows school districts the option of using response-to-intervention (RTI) methodology to identify specific learning disabilities (Fuchs & Fuchs, 2006). The use of CBM data has been incorporated into such methodology, where students are considered for special education eligibility if they continue to show a lack of adequate progress after exposure to evidence-based interventions (Fuchs & Fuchs, 2006). As a result, CBM or CBM-like measures (e.g., Dynamic Indicators of Basic Early Literacy Skills; DIBELS; Good & Kaminski, 2002) are used for eligibility decision-making purposes (Speece, Case, & Molloy, 2003). Although different RTI models have been practiced widely in American school systems, several unsolved issues remain (Kratochwill, Clements, & Kalymon, 2007). One major concern is the lack of psychometric equivalence of CBM reading (CBM-R) measures (i.e., the incomparability of those parallel probes). To date, the evidence to support the comparability of CBM-R probes has been lacking (Betts, Pickard, & Heistad, 2009; Cummings, Park, & Bauer Schaper, 2013; Stoolmiller, Biancarosa, & Fien, 2013). From a progress-monitoring perspective within RTI models, such score variability across parallel probes may cause difficulties when deciding whether to alter or modify a given student's intervention, and it affects the accuracy of high-stakes eligibility decisions. As a result, the rate of false positives and false negatives could be high (Ardoin, Christ, Morena, Cormier, & Klingbeil, 2013; Petscher, Cummings, Biancarosa, & Fien, 2013). After reviewing 171 journal articles, chapters, and instructional manuals, Ardoin et al. (2013) concluded that there is limited psychometric or empirical support for using CBM-R for progress monitoring purposes at the individual student level due to form nonequivalence. They further commented, "It is necessary to first develop CBM-R passage sets composed of equivalent level passages, procedures that allow for equating of passages to accommodate for variation in passage difficulty, or some combination of these procedures" (p. 14).

EQUATING METHODS

Initially, CBM reading passage sets were developed by randomly selecting passages from students' curricula. This method is flawed because of the considerable variability in the difficulty of texts within curricula (Hintze & Christ, 2004). Recognizing the negative effect of inconsistencies, developers of CBM passage sets (e.g., AIMSweb; DIBELS; Good & Kaminski, 2002; Howe & Shinn, 2002) used readability formulas to control passage difficulty. However, after years of study, researchers have concluded that readability formulas are poor predictors of students' oral reading fluency performance (Ardoin, Suldo, Witt, Aldrich, & McDonald, 2005; Poncy, Skinner, & Axtell, 2005).

Very recently, researchers have begun to apply equating techniques in their studies to enhance the comparability of CBM-R scores (Betts et al., 2009; Cummings et al., 2013; Stoolmiller et al., 2013). Equating is the process of adjusting for difficulty differences between test forms built to measure the same content to establish comparability of scores across forms (Kolen & Brennan, 2004). Equivalent scaling is necessary to validate the claim that the unit of measurement across passages is similar, and therefore provides evidence that scores across passages are on the same unit of measurement (Albano & Rodriguez, 2012). For instance, two passages (A and B) may be of equal difficulty with a mean of 100 WCPM but have different standard deviations (e.g., 10 words for passage A and 20 words for passage B). If two students each read 120 WCPM in passages A and B, the scores would not strictly be comparable. One student would have scored two standard deviations above the mean on passage A, while the other student would have scored only one standard deviation above the mean on passage B. Thus, an investigation of equivalence should involve an evaluation of both difficulty and scaling across passages.

Mean equating is used to adjust the distribution of scores so that the mean of one form is comparable to the mean of the other form without changing the original score scale. This is the most basic method of horizontal equating, but it is only appropriate if the standard deviations across alternate forms are similar (Christ & Hintze, 2007). The mean level of performance across multiple passages (i.e., easy and difficult) would be placed at a selected mid-point to equate performances across forms. For instance, the CBM-R mean might be scaled to 30 WCPM for all first grade CBM-R probes. If passage A had a mean of 40 WCPM, then 10 points would be taken away from each student's score because it is a relatively easy probe. However, the mean equating method might be too simplistic because it does not take into account any differences in standard deviations across the forms (Albano & Rodriguez, 2012). Nevertheless, one advantage is the scores after mean equating transformation are still authentic scores, which keep the same scale as the raw scores. This can make the use and communication of the assessment outcomes easier for the audience than defining the level of a student's performance as in relative norm-referenced scores.

Linear equating can be conceptualized as an establishment of equivalent standard scores (z-scores) for two or more different parallel forms. Moreover, linear equating can be used when the standard deviations across alternate forms are substantially different. Nevertheless, there is an important assumption that the score-distribution shapes of the different forms should be the same, or at least approximately the same (Kolen & Brennan, 2004). Linear equating adjusts scores for differences across forms in both the mean and the standard deviation, such that the rescaled scores for different forms will have the same mean and standard deviation (*SD*). However, linear equating does not eliminate nonlinear relations across forms. Also, the transformed scores after linear equating are no longer authentic scores (Christ & Hintze, 2007). Albano and Rodriguez (2012) utilized data collected in Francis et al. (2008) to demonstrate the effects of mean and linear equating transformation with the second grade DIBELS ORF probes. Given the small sample size ($N < 70$) with a partial random-groups design, the results still suggested that linear equating transformation was more preferable

than no equating for managing score variability.

Equipercentile equating is the most general and powerful method because it can accommodate any degree of linearity or nonlinearity across forms. Through the conversion of raw scores (i.e., WCPM) to equipercentile ranks, students' scores on nonequivalent forms can be compared (Kolen & Brennan, 2004). Among the three equating methods, equipercentile equating requires a relatively larger sample size to make sure percentile ranks for each point can be estimated, or it may result in a poor approximation of the equating relation due to too many unobserved points on the scale (Albano & Rodriguez, 2012). Because the required sample size and the complexity of the procedures for the mean, linear, and equipercentile equating methods increase successively, it may just be unwise to apply equipercentile equating in all conditions if simpler procedures can perform adequately (Stoolmiller et al., 2013). Since equipercentile equating was not the focus in the current study, interested readers can consult Kolen and Brennan (2004) for more details.

To examine this cost-efficiency issue and identify the most parsimonious equating method, Cummings et al. (2013) examined the three equating methods' accuracy and efficiency using DIBELS Next DORF probes from Grades 1 to 6. The mixed-model ANOVAs with repeated measures were conducted to examine the effects of the equating methods on the overall means across each grade's DORF probes. In addition, the standard error of equating (SEE; standard deviation of equated scores) information was used to index the potential effect on score variability caused by random sampling error (Kolen & Brennan, 2004). With their second grade sample for demonstration, Cummings et al. (2013) found relatively lower SEE values (compared with the mean equating method) produced by the linear equating method in the score range between 30 and 80. Further, the efficiency of the equating methods for the DORF probes was examined with the likelihood ratio test for comparing model fits based on chi-squire statistics. More detailed information regarding the testing procedure and the analyses can be found in Cummings et al. (2013). In summary, the results of their efficiency examination indicated linear equating produced the best outcomes with their sample, except for Grade 1, where mean equating was the most efficient method for 13 out of the 20 progress-monitoring probes. However, because equating procedures remove the mean differences across probes, it should have been no surprise their results showed there was no significant difference across the probe means following the equating transformation. In other words, no significant score variability was identified after the equating transformation because the equating procedures had eliminated the form nonequivalence at the group level. Although this was an important finding of their study, score variability at the individual level was left unexamined. Given the fact that the use of CBM-R progress-monitoring procedures in making high-stakes decisions is performed mainly at the individual student level, it is also very important to consider what happens to score variability at the individual level under the different equating methods. As Shapiro (2013) commented, "Certainly, the lack of evidence of the reliability and validity of decisions made from these types of individual progress-monitoring data is a potentially devastating finding to practitioners" (p. 61). In other words, the evidence for the validity of CBM-R based on group study findings should not be over generalized to its application to monitor and evaluate individual student progress

(Ardoin et al., 2013). Thus, more evidence of validation is needed for the use of CBM-R progress-monitoring procedures to inform high-stakes decision making at the individual student level. Taken together, although the current research findings suggest estimates of students' reading rate (i.e., word correctly read per minute; WCRPM) would benefit from establishing equivalent scaling to facilitate comparison of non-equivalent passages, no studies have examined the effects of the equating procedures at the individual student level.

PURPOSE OF THE STUDY

As stated in the literature, the equating methods and the trade-offs (e.g., cost-efficiency) associated with them have just begun to be investigated (Petscher et al., 2013). Since both the mean and linear equating methods were found to be more efficient at the primary grades than equipercentile equating (Cummings et al., 2013), the current study was intended to further examine and compare these two equating methods to determine their assistance with managing the equivalence of the first grade CBM-R probes. Moreover, we proposed a different analysis method from Cummings et al. (2013). We examined how much the variance in individual scores (i.e., score deviation from each individual's mean) might be reduced through statistical equating. This approach can directly validate the use of CBM-R probes to monitor progress at the individual student level, which was never examined in any previous study with a similar design of using repeated measures.

Two research questions were addressed in this study:

Research Question #1: *Does the mean or linear equating method perform better than no equating transformation for managing within-participant variances across the probes?*

Research Question #2: Does one equating method perform better than the other?

If evidence were to support the use of these less complex equating methods for making CBM-R progress-monitoring scores more comparable at the first grade level, it might serve to encourage educators and researchers to use them with greater assurance. Moreover, when CBM-R probes are more comparable after equating transformation, fewer probes may be needed to generate precise and accurate educational decisions as the magnitude of measurement errors are under better control (Hintze & Christ, 2004).

METHOD

Participants and Contexts

This study was conducted in two elementary schools located in a Midwestern state. Parental consent and student oral assent were obtained for 68 first grade students (36 females and 32 males). Thirty-one of the participants were from a rural school, in which the first grade students were 100% Caucasian. About 10% of the entire school population in the first building was eligible for discounted/free school meals. About half of their first graders (52%) participated in this study. The other 37 participants were from a small city school with an ethnically diverse student population. About 24% of the

first grade students participated in the study. The first graders in that school year were composed of Caucasian (47%), African American (29%), Hispanic (19%), and Asian/Pacific Islander (5%). About 68% of the students in the second building received free or reduced lunch.

In the first school setting, the test administrators used the same two testing rooms to administer the assessments to students individually. In the other school, the test administrators used the library and the school counselor's office. Each session lasted approximately 10 min.

Procedures and Measures

DIBELS ORF (DORF) Sixth Edition is a standardized, curriculum-based measure for indexing a reader's overall reading competence (Good & Kaminski, 2002). Reliability evidence was reported by its developers in terms of alternate-forms reliability (r = .89 to .94). The 20 progress-monitoring probes were arranged in four packets of five probes for each day of data collection. To avoid fatigue and practice effects, the 20 DORF probes were divided into four sets of five probes and were intended to be administered in a counterbalanced order in 4 days. However, a perfect counterbalanced design (i.e., having 17 participants for each probe set on each testing day) was not obtained during administration. Table 1 shows the exact numbers of participants administered each probe set on each data collection day. Five probes in each set were randomly administered to each participant to avoid order effects. For example, on Day 1 each of the 19 participants might receive a different order with the probe Set 1 such as 3, 1, 4, 5, 2, or 4, 1, 2, 3, 5. On Day 2, each of the other 15 participants might receive a different randomly assigned order such as 5, 4, 1, 2, 3, or 2, 4, 1, 3, 5. Therefore, the administration orders varied for each participant.

During assessment, the test administrator placed the probe in front of the student and read scripted instructions to students prior to reading the probes informing them that they would be reading aloud, explaining where to start reading, and encouraging them to do their best. Each participant read the 20 progress-monitoring DORF probes during a 1-week interval at the end of the school year. According to Kolen and Brennan (2004), the use of such repeated measures across all forms with a single group of subjects can better control random errors than using a random-groups design. In addition, each participant was informed that he or she would be allowed to choose a sticker from the examiner as a reward at the end of each day's assessment. The participants were also told that they would receive an ice-cream gift card when they finished all the probes at the end of the study. Using incentives could ensure that any changes in student performance on probes within and across days were a function of changes in the difficulty of the probes and not changes in student motivation. At the end of each data-collection day, the researcher according to the standardized directions completed the scoring for scoring in the technical manual of DIBELS (Good & Kaminski, 2002).

Inter-Administrator Agreement

There were four test administrators and all were trained in DORF administration procedures. One administrator was a senior school psychologist who had 15 years' experience working for the local educational agency

Table 1. *Numbers of Participants in the Four Groups of Probe Sets on Each Testing Day*

	DAY 1	DAY 2	DAY 3	DAY 4
Set 1 (Probe 1~5)	19	15	17	17
Set 2 (Probe 6~10)	20	18	14	16
Set 3 (Probe 11~15)	17	19	19	13
Set 4 (Probe 16~20)	12	16	18	22
Total Participants	68	68	68	68

(LEA). She received DORF training provided by the LEA in which she was employed. The other three test administrators were doctorate school psychology program students from the author's institution. Directions for administration were reviewed by the researcher to them, including the standardized directions verbatim, the coding system (i.e., a slash for an incorrect response, a bracket after the last word provided at the end of 1 min), and other rules (e.g., discontinue rule, hesitating or struggling with words). After review of the standardized administration procedures, a randomly selected first grade DORF benchmark probe was used for practice with each administrator. During the training sessions, the researcher pretended to be a beginning reader and made common mistakes (e.g., omission, commission, repetition, jumping through lines) to familiarize the administrators with the recoding procedures. Each training session was about a half hour and on an individual basis with the administrators.

Inter-administrator agreement was examined to ascertain the degree of scoring accuracy. The researcher and one of the four test administrators served as the primary test providers and the other three administrators served as independent recorders. While the primary administrator was administering the DORF probes, an independent recorder was recording the data in a separate examiner booklet using the DORF coding system. Inter-rater agreement was assessed for about 20% of the assessment data on a word-by-word basis by comparing each word the test administrators recorded as correct or incorrect to each word the independent recorder scored as correct or incorrect. The number of agreements (correct and incorrect) between the administrator and the independent recorder was divided by the total number of words and multiplied by 100 to obtain a percentage (House, House, & Campbell, 1981). The results indicated that the judgments made by the test administrators on each probe had high inter-rater agreement with an average agreement of 98%.

Equating Transformation

Words correctly read per minute (WCPM) was calculated using the criteria in the DIBELS manual (Good & Kaminski, 2002). Then, we applied mean equating and linear equating to equate passages to the scale of the first probe (*the Ant Hill*), which showed the smallest value of standard deviation (*SD*) with the current sample. Because linear equating modifies each probe's variance to match the reference probe, it was important to select the

passage with the smallest variance in its score distribution as the reference. In this way the rescaled variance was kept to a minimum. Also, Albano and Rodriguez (2012) suggested the reference passage should be about average difficulty to ensure the overlap between score distributions across passages. The average score of *the Ant Hill* probe with the current sample was about 3 WCPM above the overall mean of the 20 DORF probes, which addressed this suggestion.

Data Analyses

In a repeated measure ANOVA, the total variation, SS_{Total} can be partitioned into $SS_{Between.Persons}$ and $SS_{Within.Persons}$. In an experimental design, $SS_{Between.Persons}$ is a function of differences between the means of the persons who receive treatments and $SS_{Within.Persons}$ is a function of the pooled variation within the individual persons across the treatments. Different to Cummings et al. (2013), who examined the overall treatment effects resulting from the equating transformation (i.e., $SS_{Between.Persons}$), we examined how much score variability could be reduced (or under control) in each participant by the selected equating methods (i.e., $SS_{Within.Persons}$). In other words, less score variability might be observed in a participant's graph of scores. In everyday practice, educators are usually more interested in the within-participant variation that could interfere with the accuracy of decision making (e.g., concluding that changes in progress-monitoring scores reflect a student's response to instruction). The within-participant variation can be partitioned into the effect of treatment variation (in this case the probes) and residual variation. In this study, the residual variation could be due to participant-probe interaction, temporary performance fluctuation, and other uncontrolled residual sources other than the effects caused by the equating procedures. To determine the within-participant variance, we calculated the variation within each participant. The formula was $SS_{w.person\ i} = \sum(Y_{ik} - M_{pi})^2$. This is the sum of the squared deviations of the scores for person i away from the mean for person i. As mentioned earlier, this reflects score variability associated with the progress-monitoring probes and residual errors. The degrees of freedom (*df*) for the within-participant variation is k - 1. Dividing each obtained within-participant variation by its degrees of freedom, yielded the within-participant variances for the 68 participants in the study. Thus, instead of examining overall treatment differences in the equated scores at the group level as was done in Cummings et al. (2013), we used a repeated measures ANOVA to determine whether there was a significant difference among the within-participant variances under the three conditions: no equating, mean equating, and linear equating.

As a follow up to the overall repeated measures ANOVA, we conducted planned multiple comparisons using the Bonferroni procedure to determine if there were significant differences between the means of within-participant variances under the three equating condition. The level of significance was set at $\alpha = .05$ for the repeated measures ANOVA and the level of significance set at a family-wise error rate of .05 for the multiple comparisons.

Besides ANOVA, the SEE allowed further analysis of the accuracy of the two equating methods (Kolen & Brennan, 2004). The SEE applied to each DORF probe was calculated through the bootstrap method over 500 replications of the two equating methods for comparisons.

Similar to Cummings et al. (2013), the obtained SEE results in the current study should be interpreted with reservation due to the relatively small sample.

Further, in order to demonstrate the difference in score transformation between the three equating conditions, the data of the lowest performing reader with our sample was used for this purpose. The reason for choosing this dysfluent reader is that DORF progress-monitoring probes are usually used for tracking progress of those students who score below benchmarks. The benchmark of DIBELS DORF score for the end of the first grade is 40 WCPM (Good & Kaminski, 2002). The selected student, who read an average of approximately 22 WCPM in this study, fell much below the benchmark and is suitable for the propose of this demonstration. This student's raw and rescaled scores were graphed to allow visual analyses of the magnitude of the DORF scores fluctuating across probes at the individual student level and the effects of the two equating methods on passage equivalence.

RESULTS

Performance before Equating

The current study applied a single-group equating design to control random errors in CBM-R data collection. The results of descriptive statistics indicated that the average raw scores (WCPM) across probes encompassed a range from a low of 69 WCPM on probe #16 to 88 WCPM on probe #18. The range of variability indexed by the standard deviation (*SD*) was from 31.5 to 40.3. The score distributions of the 20 DORF probes were close to a normal distribution. The alternate form correlations between the 20 probes were quite high, .89 to .97, which were consistent with previous findings (e.g., Betts et al., 2009; Stoolmiller et al., 2013).

Effects of Equating Methods

Table 2 shows the means and standards of the within-participant variances under the three probe equating conditions. Preliminary analysis indicated the repeated measures ANOVA assumption of sphericity was not met, so the results were reported for the lower-bound conservative test. The effect of the equating methods on the within-participant variances was statistically significant, $F(1, 67) = 130.23$, $p < .001$. Following the advice and formulas provided by Tabachnick and Fidell (2007, p. 290) for estimating measures of effect size when the sphericity assumption is violated, partial η^2 was computed to be .66 and the lower-bound value for η^2 was computed to be .29. The finding indicated a significant difference among the within-participant variances under the three treatment conditions that explained between 29% to 66% of the within-subject variances. Multiple comparisons using a Bonferroni procedure with separate error terms revealed the within-participant variances were significantly lower ($p < .01$) under the linear equating method ($M = 65.35$) than the mean equating method ($M = 90.64$, $d = -0.30$) or the no equating condition ($M = 129.53$, $d = -0.77$). In addition, the within-participant variances were significantly lower ($p < .01$) after the mean equating transformation than the no equating correction ($d = -0.47$). Due to violation of the sphericity assumption, we based our calculations of the Cohen's *d* effect sizes on the standard deviation of the no equating control condition as recommended by Cohen (1988), which reduced the effect size estimates in this

Table 2. *Descriptive Statistics for the Within-Participant Variances under the Three Equating Conditions*

	M	*SD*	95% CI for *M*
No Equating	129.53	83.00	[109.44, 149.62]
Mean Equating	90.64	65.31	[74.84, 106.45]
Linear Equating	65.35	47.36	[53.89, 76.81]

case. Importantly, the results indicated the mean equating method and the linear equating method both reduced the within-participant variances across the probes when compared to the no equating condition. Using the no equating method as the base, the results indicated an average reduction of 49.5% in the within-subject variances following linear equating and a 30% reduction in the within-subject variances following the mean equating method. Thus, the use of the equating methods made a substantial difference to the amount of fluctuation in individual performances across the probes. The effects of the equating methods on progress-monitoring decision making at the individual level were further demonstrated in the following section via visual analyses of raw and rescaled scores.

As to the SEE analysis of the two equating methods, the result of using the progress-monitoring Probe 2 as an example is presented in Figure 1 for demonstration. The estimated SEE of the mean equating method was about 5.64 and was consistent across all score levels in the distribution. The estimated SEE values associated with the linear equating method varied at each score point and were relatively lower when there were more frequent data points presented. Specifically, the estimated SEE values with the linear equating were lower than those with the mean equating transformation through the scores ranging from 35 to 89 WCPM. In other words, without frequently observed cases at the two ends of the score distribution, the estimated effect of the linear equating might not be as strong as the mean equating method due to sampling error. These findings regarding the SEE patterns with the two equating transformation were similar to what was found in Cummings et al. (2013) with their second grade sample. Also, these SEE results were consistent across the other DORF probes (i.e., Probe 3 to 20) in the current study.

Visual Demonstration

To demonstrate probe effects and potential benefits of the two equating procedures at the individual student level, Figure 2 and Figure 3 were used to show the raw and rescaled scores of the first grade DORF with the mean and linear equating methods for the lowest performing participant, Clifford. First, a significant magnitude of passage effects across the 20 probes was observed in Clifford's raw scores (up to a difference of 25 WCPM). With visual analyses, the score variability was substantially reduced with both equating methods. In general, the linear equating functioned better in reducing score variability. Specifically, the mean equating resulted in a maximum difference (i.e., highest score - lowest score)

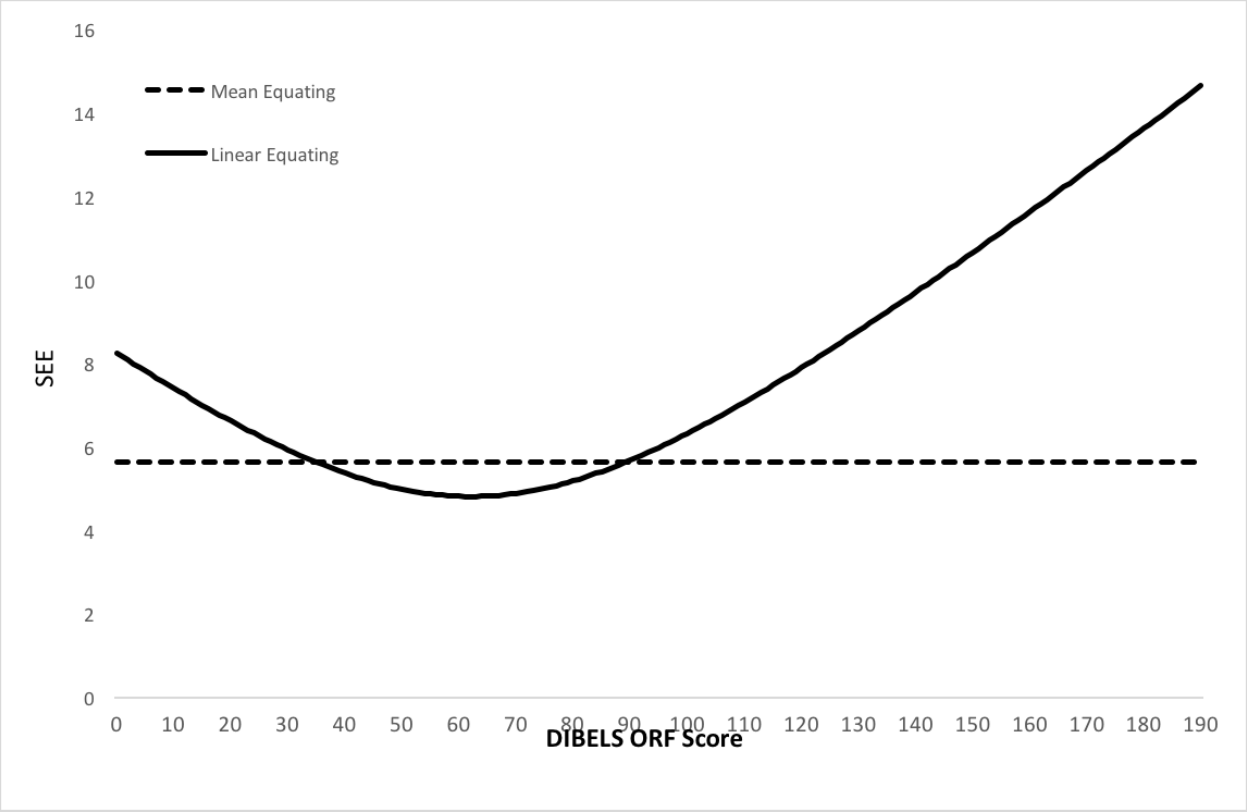

Figure 1. The standard errors of the mean and linear equating methods (SEEs) of the first grade DORF progress-monitoring Probe 2.

of 18 WCPM between the 20 progress-monitoring probes. In contrast, the linear equating showed a maximum difference of only 14 WCPM. Thus, the visual inspection confirms the ANOVA results regarding the effects of the equating methods on within-participant variance and demonstrates that the linear equating method had the most desirable outcome for managing score variability when examining individual performances for progress-monitoring decisions.

DISCUSSION

IDEIA (2004) allows using RTI data as a part of the procedure of identifying specific learning disabilities. In practice, however, no special education eligibility decisions are made for a group of referred students but only for individuals. Thus, it is important to study the use of test scores for this specific purpose (Ardoin et al., 2013; Messick, 1989). To our knowledge, the current study is the first examining the effects of CBM-R score variability within participants. The linear equating method in this study resulted in a significantly smaller within-participant variance than the mean equating method or no equating transformation. This result was consistent with the findings in previous studies that any equating methods outperformed no equating in terms of reducing score

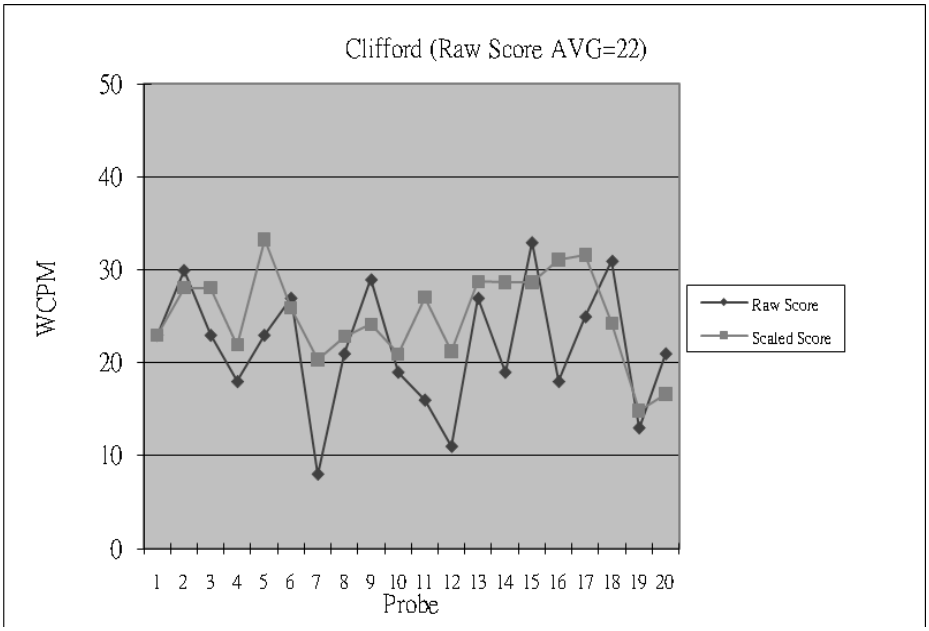

Figure 2. The comparison between the raw and rescaled scores with mean equating on the first grade DORF measures with Clifford.

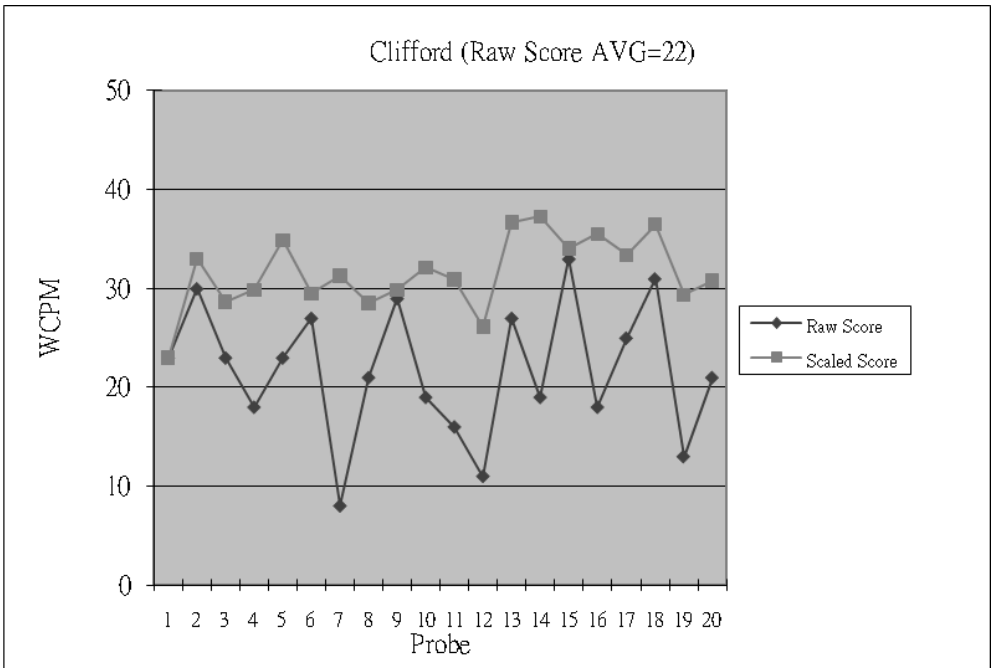

Figure 3. The comparison between the raw and rescaled scores with linear equating on the first grade DORF measures with Clifford.

variability across probes (Albano & Rodriguez, 2012; Cummings et al., 2013). However, different from what Cummings et al. (2013) had found, our finding indicated that the linear equating method rather than mean equating was more favored with our first grade sample. One possible explanation is that their data were collected in the middle of the first grade while ours were collected at the end. Therefore, our participants' scores might result in an approximate normal distribution with less impact by floor effects.

Different from the comparisons of overall equating effects using ANOVA, the analyses of the estimated SEE values painted a slightly different picture with the present sample. The linear equating method was more favored for score transformation only for those scores falling between 35 and 89 WCPM because beyond that score range, the low coverage of the data points may create more error than it can remove. Since DORF progress-monitoring probes are usually used with those low-performing students who are at the lower end of the distribution, this finding reemphasized the importance of equating methods selection as raised in the previous studies with relatively small sample sizes applied (e.g., Cummings et al., 2013; Francis et al., 2008). In sum, assuming the SEE is well controlled with an appropriate sample size (i.e., having enough data points at each score level), our finding suggests the linear equating method will be a better option to produce more interchangeable CBM-R scores when progress monitoring first grade readers. This significant reduction of score variability could help school teachers and school psychologists make more accurate decisions about their students' progress.

Probe Nonequivalence

The key to success of an RTI model is the availability of measures suited for frequent progress monitoring to track student performance over time. Specifically, the reading passages need to function as parallel forms so educators can conclude that changes in the scores on the measures actually reflect changes in student performance, not measurement errors (Hintze & Christ, 2004). Although the developers of the DIBELS ORF measures made significant efforts to control for passage differences by using multiple readability formulas, significant differences in raw scores across the DORF probes were identified in the current study. This has also been shown in prior research (Betts et al., 2009; Cumming et al., 2013; Stoolmiller et al., 2013). As concluded in Cumming et al. (2013), "With this level of passage variability, school teams will struggle with identifying the extent to which student gains or losses in reading performance are due to true changes in reading skill, behavioral problems (e.g., lack of student motivation), or passage difficulty" (p. 103). In the present study, the score variability within each participant could not be fully eliminated by the equating procedures. Even after the most desirable equating transformation (i.e., linear equating), the maximum difference was as high as 14 WCPM. Yet, the literature suggests for children in general education, realistic growth expectation for first graders is only 2 WCPM per week (Deno, Fuchs, Marston, & Shin, 2001; Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993). Thus, the score variability is significantly larger than the expected weekly growth, which makes it difficult to develop reliable and valid decision rule for evaluating individual students' response to intervention. According to the modern

concept of validity which places a heavy emphasis on how a test is used (Messick, 1989), this observed score variability across parallel forms would serve as a psychometric threat to the use of CBM-R results for making educational decisions at the individual student level.

In fact, equating procedures, regardless of the type, can only take care of score variability as reflected in an overall effect (e.g., Cumming et al., 2013). Each rescaled probe shared the same mean after the mean equating transformation or the same mean and standard deviation after the linear equating transformation. However, the individual score variability across probes remained. Although recent equating studies (Cumming et al., 2013; Stoolmiller et al., 2013) have found a positive impact to the application of equating methods on reducing CBM-R score variability, such results of comparability based on group designs should not be automatically transferred to making individual decisions without caution. This finding extended our understanding of the probe effects at the individual student level. To address this unsolved issue regarding the nonequivalence in CBM-R probes, other statistical methods such as generalizability (G) theory may be used with the equating techniques as a combination to improve the validity of individual decision making (Fan & Hansmann, 2015; Petscher et al, 2013). Future study may also consider examining the effect of a combination of different statistical procedures as to managing score variability with CBM-R.

Practical Implications

To ensure the accuracy of using equating transformation, test developers (e.g., AIMSweb; DIBELS NEXT; Good & Kaminski, 2011; Howe & Shinn, 2002) may consider the significance of this finding when developing the next versions of CBM-R assessments to improve the precision of progress-monitoring decisions. In operation, it seems more reasonable for those enterprises to recruit a large sample of students across all grade levels to develop psychometrically sound parameters for cross validation before the dissemination of their CBM-R products. For example, the test developers ought to develop and include score conversion tables in their technical manuals for potential users to transform raw scores into equated scores to help make more accurate decisions. As suggested by Stoolmiller et al. (2013), such instrument development activities (e.g., examining the effects of different equating methods to other non-analytic sample) should be conducted by test developers with advanced knowledge of measurement and statistical equating methodologies.

As to practicability, Nitko (1996) named the practicality features one of the eight facets to validity evidence of test use. A test (or statistical method) may not result in adequate outcomes if its operation is not perceived as cost-efficient by its users. In short, they may not even consider using it. Thus, not only theoretical but also practical factors for each specific use of equating procedures need to be thoroughly considered such as the required sample size, training and efforts for operation, and acceptability/understandability of the transformed scores by its potential audience (Albano & Rodriguez, 2012; Stoolmiller et al., 2013). Therefore, debates between using absolute or relative norm-referenced scores to describe a student's authentic performance remain (Cummings et al., 2013). In this study, the result of the pairwise comparisons suggested that the linear

equating method outperformed the mean equating condition and the mean equating method outperformed no equating transformation at the first grade level. A simple implication is that having either mean or linear equating transformation is better than no equating correction. However, when choosing between mean or linear equating methods to manage nonequivalence in CBM-R scores, linear equating may not always be favored even with its superior outcome in the current study because the absolute scores (WCPM) generated by the mean equating method would be more understandable and acceptable by the general population to describe a student's authentic performance than using a relative score (after linear equating transformation).

Limitations and Future Research Directions

The results of the present study are specific to the sample and the measures described. Several limitations must be acknowledged. First, the current study included only 68 first grade students. A larger sample of students is usually preferred for using an equating procedure (Kolen & Brennan, 2004). Future studies should include larger sample sizes to reduce sampling error and consider examining the benchmark probes so each progress-monitoring score can be directly equated back to the benchmark result. However, the primary purpose of the present study was to compare the effects of variance reduction at the individual level under the three equating conditions rather than establishing parameters based on the current first grade sample to allow future application with another sample. Thus, due to the different purposes of data use, not having ideal sample sizes for cross validation would be unlikely to affect the importance of the present study. This study showed equating methods made a difference to first grade DORF score variability at the individual student level, which has not yet been studied and addressed in the current literature. Our results merit the attention of both researchers and educators regarding the validation of the use of CBM-R with statistical equating at the individual student level. Second, results are limited by examining only one grade level, as well as the selection of participants from two Midwest elementary schools. The current results should not be automatically generalized to other subpopulations or other CBM-R measures without further replications of the findings. Third, a distributive model of treatment acceptability (Carter, 2008) may guide future research to investigate the practicability features of different equating methods. This model comprises three facets: consumer acceptability, consultant acceptability, and societal acceptability. School psychologists are typically consultants who have training and experience to implement and monitor the use of equating procedures in educational settings. However, teachers' and other professionals' psychometric knowledge and previous experience should also be taken into account. Without adequate buy-in from consumers, the practicability evidence for the proposed equating methods might be weak (Nitko, 1996). In other words, the gap between scientific findings and the real world practice remains. It is recommended that formal or informal methods used to assess acceptability may include rating scales or interview.

CONCLUSION

Since CBM procedures play an increasingly important role

in making high-stakes educational decisions, it is important to understand their limitations when using their outcomes for different assessment purposes. The current study examined the effects of mean and linear equating methods for managing score variability within individual first graders. The results indicated that the form effects can be effectively controlled by the two equating methods and the comparability of the DORF scores was significantly improved. However, the score variability was still not negligible at the individual student level. Future research should further investigate the effects of those equating methods for tracking individual progress-monitoring data to directly address the validity issue (i.e., how a test is really used) by considering the purposes and interpretation of using DORF outcomes at RTI tiers (Ardoin et al., 2013). This is the primary implication of the present study.

REFERENCES

Albano, A. D., & Rodriguez, M. C. (2012). Statistical equating with measures of oral reading fluency. *Journal of School Psychology*, *50*(1), 43-59.

Ardoin, S. R, Christ, T. J., Morena, L. S., Cormier, D. C., & Klingbeil, D. A. (2013). A systematic review and summarization of the recommendations and research surrounding curriculum-based measurement of oral reading fluency (CBM-R) decision rules. *Journal of School Psychology, 51*, 1-18. doi:10.1016/j.jsp.2012.09.004

Ardoin, S. P., Suldo, S. M., Witt, J. C., Aldrich, S., & McDonald, E. (2005). Accuracy of readability estimates predictions of CBM performance. *School Psychology Quarterly,20*, 1−22.

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

Carter, S. L. (2008). A distributive model of treatment acceptability. *Education and Training in Developmental Disabilities, 43*(4), 411–420.

Christ, T. J., & Hintze, J. M. (2007). Psychometric considerations of reliability when evaluating response to intervention. In S. R. Jimmerson, A. M. VanDerHeyden, & M. K. Burns (Eds.), *Response to intervention handbook* (pp. 93-105). New York: Springer.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cummings, K. D., Park, Y., & Bauer Schaper, H. A. (2013). Form effects on DIBELS Next oral reading fluency progress-monitoring passages. *Assessment for Effective Intervention, 38*(2), 91-104. doi: 10.1177/1534508412447010

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219−232.

Deno, S. L., Fuchs, L. S., Marston, D., & Shin, J. (2001). Using curriculum-based measurement to establish growth standards for students with learning disabilities. *School Psychology Review, 30*, 507–524.

Fan, C., & Hansmann, P. R. (2015). Applying generalizability theory for making quantitative RTI progress-monitoring decisions. *Assessment for Effective Intervention, 40*(4), 205-215. doi: 10.1177/1534508415573299

Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. R. (2008). Form effects on

the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology, 46*, 315−342.

Fuchs, D., & Fuchs, L. S. (2006). Introduction to response to intervention: What, why, and how valid is it? *Reading Research Quarterly, 41*, 92-99.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Walz, L., & Germann, G. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review, 22,* 27-48.

Good, R. H., & Kaminski, R. A. (Eds.). (2002). *Dynamic indicators of basic early literacy skills* (6th ed.). Eugene, OR: Institute for the Development of Education Achievement. Available: http://dibels.uoregon.edu

Good, R.H., & Kaminski, R.A. (2011). DIBELS Next assessment manual. Eugene, OR: Dynamic Measurement Group. Available: https://dibels.org/

Hintze, J. M., & Christ, T. J. (2004). An examination of variability as a function of passage variance in CBM progress monitoring. *School Psychology Review, 33*, 204-217.

House, A. E., House, B. G., & Campbell, M. B. (1981). Measures of interobserver agreement: Calculation formula and distribution effect. *Journal of Behavioral Assessment, 3*, 37-57.

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

Individuals with Disabilities Education Improvement Act, 20 U.S.C 1400 et seq. (2004).

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Kratochwill, T. R., Clements, M. A., & Kalymon, K. M. (2007). Response to intervention: Conceptual and methodological issues in implementation. In S. R. Jimerson, M. K. Burns, & A. M. Van Der Heyden (Eds.), *The handbook of response to intervention: The science and practice of assessment and intervention* (pp. 25-52). New York, NY: Springer.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Nitko, A. J. (1996). *Educational assessment of students*. Englewood Cliffs, NJ: Merrill/Prentice Hall.

Petscher, Y., Cummings, K. D., Biancarosa, G., & Fien, H. (2013). Advanced (measurement) applications of curriculum-based measurement in reading. *Assessment for Effective Intervention*, *38* (2), 71-75. doi: 10.1177/1534508412461434

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*(4), 326–338. doi: 10.1177/073428290502300403

Shapiro. E. S. (2013). Commentary on progress monitoring with CBM-R and decision making: Problems found and looking for solution. *Journal of School Psychology, 51*, 59-66. doi: 10.1016/j.jsp.2012.11.003

Shinn, M. (Ed.). (1998). *Advanced applications of curriculum-based measurement.* New York, NY: Guilford.

Speece, D. L., Case, L. P., & Molloy, D. E. (2003). Responsiveness to general education instruction as the first gate to learning disabilities identification. *Learning Disabilities Research & Practice, 18*, 147-156.

Stoolmiller, M., Biancarosa, G., & Fien, H. (2013). Measurement properties of DIBELS oral reading fluency in grade 2: Implications for equating studies. *Assessment for Effective Intervention*, *38*(2), 76-90. doi: 10.1177/1534508412456729

Tabachnick, B. G., & Fidell, L. S. (2007). *Experimental design using ANOVA*. Belmont, CA: Thomson.