ELSEVIER

Journal of School Psychology 48 (2010) 39 – 52

Journal of School Psychology

Curriculum-based measurement of math problem solving: A methodology and rationale for establishing equivalence of scores

Marjorie Montague [a,*], Randall D. Penfield [a], Craig Enders [b], Jia Huang [a]

[a] *University of Miami*
[b] *Arizona State University*

Received 13 August 2009; accepted 18 August 2009

Abstract

The purpose of this article is to discuss curriculum-based measurement (CBM) as it is currently utilized in research and practice and to propose a new approach for developing measures to monitor the academic progress of students longitudinally. To accomplish this, we first describe CBM and provide several exemplars of CBM in reading and mathematics. Then, we present the research context for developing a set of seven curriculum-based measures for monitoring student progress in math problem solving. The rationale for and advantages of using statistical equating methodology are discussed. Details of the methodology as it was applied to the development of these math problem solving measures are provided.

Keywords: Curriculum-based measurement; Methodology; Assessment

The primary purpose of this article is to describe the development of a set of seven curriculum-based measures for monitoring students' progress in math problem solving. Our measures were developed specifically for a research project investigating the effects of an intervention to improve math problem solving for middle school students. First, a brief

* Corresponding author.
E-mail address: mmontague@aol.com (M. Montague).
ACTION EDITOR: Edward J. Daly, III.

0022-4405/$ - see front matter © 2009 Society for the Study of School Psychology. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.jsp.2009.08.002

history of curriculum-based measurement (CBM) and a review of the CBM literature are provided. Second, the research context for developing our measures is discussed. Third, the rationale and methodology for statistically equating the measures are described. Finally, the results of the analyses supporting the need for using equated ability estimates for measuring students' academic progress over time are presented.

CBM history and research

Deno and Mirkin (1977) are credited with originating CBM as a method to monitor students' academic progress. CBM was developed as a measurement system to test the effectiveness of a special education intervention model (i.e., data-based program modification) by obtaining valid and reliable repeated measures of students' academic performance in order to evaluate and improve instruction (Deno, 1985; Deno, Fuchs, Marston, & Shin, 2001). In educational decision-making, CBM is used for screening, identifying, and referring students at risk for academic failure; gauging students' responsiveness to interventions; evaluating the effects of interventions; making instructional decisions; and, most recently, predicting students' achievement on high-stakes assessments (Fuchs & Deno, 1992; Deno, 2003; Fuchs, 2004; Fuchs, Fuchs, & Courey, 2005; Madelaine & Wheldall, 1999). With the reauthorization of the Individuals with Disabilities Education Improvement Act (IDEIA) in 2004, Response to Intervention (RtI) was included as an alternative method to identify students with learning disabilities (LD). RtI is predicated on the idea that all children receive high quality instruction in general education classrooms. Children who do not make adequate progress despite high quality instruction as determined by ongoing assessment are then provided with increasingly intense, multi-tiered interventions that may eventually result in special education placement. Consequently, RtI models require progress monitoring or frequent assessment of student performance to make appropriate instructional decisions for children (Busch & Reschly, 2007; Fuchs & Fuchs, 2006; Speece, Case, & Molloy, 2003; Wallace, Espin, McMaster, Deno, & Foegen, 2007).

Deno and Fuchs (1987) noted the importance of knowing what to measure, how to measure it, and also how to use the resulting data for making educational decisions. They underscored three criteria that must be met if CBM is to be viewed as a credible measurement system. That is, CBM must be technically adequate, able to determine instructional effectiveness, and logistically feasible. The selection of what to be measured has been viewed as the most important concern in developing CBM because the targeted performance needs to be responsive to the effectiveness of instruction through repeated measurement. Therefore, what is measured needs to be specified (i.e., the task) and indicators for growth with respect to the task should be determined before developing CBM. Thus, CBM must efficiently measure performance in a specific area and result in reliable and valid data that document a student's growth over time. Deno (2003) noted that this characteristic distinguishes CBM and continued to say that "repeated observations of performance are structured so that students respond to different but *equivalent* (our italics) stimulus materials that are drawn from the same general source" (p. 185). Unfortunately, CBM research generally has failed to provide empirical evidence of the equivalency of the different stimulus measures that are used to monitor student progress. This is a major concern given the important role that CBM promises to play in the context of RtI models

that will be the mechanism for making critical educational decisions for students (e.g., special education placement). Across academic domains, it is essential that alternate forms used in CBM are not only technically sound (i.e., reliable and valid) but also yield scores that are equivalent. Although there are numerous CBM studies focusing on reading, particularly early reading skills, we found only two studies that addressed equivalence of alternate forms. Both studies focused on CBM of oral reading, a technique referred to as CBM-R that is commonly used to measure reading growth of children in elementary school.

CBM-R research

In the first study, Betts, Pickart, and Heistad (2009) emphasized the importance of obtaining equivalent scores from different forms so that decisions are based on actual evidence and not on the "fluctuations of the measurement properties" that characterize the instruments. They challenged the claims that passages used in CBM of oral reading fluency (ORF) are equivalent by examining the use of readability statistics as a means for equating alternate forms of passages for between-grade and within-grade assessments (Betts et al., 2009). As hypothesized, they found that although readability statistics may provide a rough estimate of between-grade differences, raw CBM scores (i.e., words read correctly per minute) typically lack equivalence and therefore should not be used for assessing within-grade growth.

In the second study, Christ and Ardoin (2009) studied four procedures for evaluating and selecting passages for CBM of oral reading including three fairly traditional methods (i.e., random sampling, readability formulas, mean level of performance evaluation) and Euclidean Distance evaluation, a novel statistical approach that considers differences across students on repeated measurements. Results corroborated that random sampling from curricular or curricular-type materials, the most common method for selecting passages for assessing ORF, does not ensure equivalence of passages. Similar to Betts et al. (2009), these authors concluded that CBM based on readability formulas was a weak method for assessing ORF as the passages frequently varied considerably in difficulty level. In contrast, they found the other two approaches were more promising but still may be unreliable when using raw scores to estimate level of performance. The point of reviewing these recent examinations of CBM in reading is to underline the importance of establishing equivalence of alternate forms for monitoring academic performance. Betts et al. (2009) suggested that modern item response theory (IRT) has potential for moving the CBM field forward by, in the case of ORF, placing passages on a scale that would statistically equate them so that scores across passages would be comparable. Compared with reading, CBM research in mathematics (CBM-M) has been minimal.

CBM-M research

Foegen, Jiban, and Deno (2007) at the Research Institute on Progress Monitoring found only 32 studies that addressed CBM for mathematics. None of the studies addressed the problem of establishing parallel forms that would ensure equivalence of scores across multiple assessments. Technical adequacy was usually indicated by reporting reliability and validity data including internal consistency, test–retest reliability, alternate form reliability, and concurrent and predictive criterion validity. Mathematics is a challenging academic area to assess as it has multiple strands or topics and within each strand (e.g., geometry)

there are specific and unique concepts to be learned and applied. Most of the CBM research in mathematics has focused on computational fluency in elementary school math rather than early mathematics development (pre-school through grade 1) or secondary school mathematics.

For example, in elementary mathematics, the Monitoring Basic Skills Progress measures (MBSP) have been widely utilized and validated through a series of studies for over 20 years (e.g., Fuchs, Fuchs, Hamelett, Stecker, 1990; Fuchs, Fuchs, Hamelett, 1989; Fuchs, Fuchs, Hamelett, Allinder, 1989; Fuchs, Fuchs, Hamelett, Stecker, 1991; Fuchs et al., 1994; Shapiro, Edwards, & Zigmond, 2005). The MBSP consists of measures of computation and concepts/applications available in a software package. The computation measure, 30 parallel forms for each grade from 1 through 6, was created by drawing problems representing a proportional sampling of computation skills from the Tennessee state curriculum. The concepts/application measure was similarly developed for grades 2 through 6. Although technical adequacy (e.g., alternate form reliability) for the *parallel forms* is reported, there is no evidence to suggest that these forms are indeed parallel and that scores from these measures are statistically equivalent and thus comparable. With respect to CBM-M in middle school, Foegen and Deno (2001) examined CBM with 100 students in grades six through eight using two alternate forms of each of the measures they developed: (a) the Basic Math Operations Task, (b) the Basic Estimation Task, and (c) the Modified Estimation Task. Helwig and Tindal (2002) developed a 15-item general outcome measure of mathematics and administered four alternate forms of this measure to 117 8th-grade students across an academic year. The measures in both studies seemed technically adequate, valid when compared with a criterion measure, and sensitive to students' growth. However, there was no evidence of equivalence between or among the alternate measures. Our CBM-M research focuses on the development of seven alternate forms that produce statistically equivalent scores to measure growth in math problem solving for middle school students.

Research context for CBM-M development

The research context for the development of these curriculum-based measures is a federally funded intervention study (2007–2010) to improve math problem solving for middle school students. The purpose of this three-year study is to test the efficacy of *Solve It!* (Montague, 2003), an intervention designed to teach students with math difficulties how to understand, analyze, solve, and evaluate mathematical problems by developing the processes and strategies that effective problem solvers use. One of the primary research questions focused on the effects of Solve *It!* on growth in math problem solving over a school year as measured by curriculum-based measures of math problem solving. A pilot study with 312 students was conducted during the first year of the project. Four schools matched on state assessment performance level (2 high-performing, 2 low-performing) and socio-economic status were recruited from the Miami-Dade County Public Schools (M-DCPS) to participate in the study. One school from each matched pair was randomly assigned to the intervention condition, and the remaining school served as a comparison. Two general education math teachers (one grade 7 and one grade 8) from each school were nominated by a school administrator to participate. Participating teachers attended a *Solve*

It! professional development workshop prior to implementing *Solve It!* in their classes that included low-achieving students and students with LD.

The math word problems for the alternate forms of the curriculum-based measures were selected from the *Solve It!* manual (Montague, 2003) and consisted of one-, two-, and three-step word problems like the following:

A store sells shirts for $13.50 each. On Saturday, it sold 93 shirts. This was 26 more than it had sold on Friday. How much did the store charge for all the shirts sold on both days?

To solve these problems, students needed knowledge of the four basic operations using whole numbers and decimals. They did not need to know specific formulas or have unique mathematical knowledge to solve the problems.

Generally speaking, test developers endeavor to construct equivalent forms of the same measure in content and difficulty but, despite good intentions, the forms typically have inevitable differences in difficulty that may prevent any meaningful interpretation when monitoring student progress. That is, because raw scores do not have intrinsic normative meaning, their use in determining progress over time can easily result in a misinterpretation of a student's underlying ability particularly when that student takes a more difficult form of the test than that taken by another student. Equating is one method that can remedy the problems associated with differences in difficulty levels of alternate test forms. Equating is a statistical procedure that transforms raw scores into scores that are comparable across alternate forms of a test. This procedure offers several theoretical and practical advantages as it places raw scores on a comparable metric so that they can be used interchangeably. It also attempts to avoid misinterpretations because any differences in difficulty of alternate forms are statistically controlled. Examinees have the same estimates of ability regardless of measurement error because the estimates are all on a common metric and, therefore, independent of the test forms. Equating methods lead to more reliable and interpretable results.

Equating methodology

The alternate forms for assessing progress in math problem solving developed for our intervention research were constructed in the following manner. Initial item selection resulted in an item bank of 30 items. Based on these 30 items, seven assessments were developed such that each assessment contained 10 of the 30 items. Each CBM form consisted of 2 one-step, 6 two-step, and 2 three-step math word problems and was administered at one of the seven time points. Because there were only 30 items, each item appeared on more than one of the seven assessments. The goal of the equating design was to create the seven assessments in a manner that would permit the analysis of growth in math problem solving proficiency across the seven time points.

One option for creating the seven assessments was to divide the pool of 30 items into three roughly parallel forms (e.g., Forms A, B, and C) for which each had a unique set of ten items. The assessments then would be administered sequentially across the seven time points such that Form A was given at Time 1, Form B was given at Time 2, Form C was given at Time 3, Form A was given a second time at Time 4, Form B was given a second time at Time 5, Form C was given a second time at Time 6, and Form A was given a third time at Time 7. This assessment plan, although conceptually simple, is fraught with limitations for a valid

assessment of growth. The first limitation is that each form was given multiple times. Because students would most likely have a memory of a specific form, their performance would be affected. The second limitation is that the three forms do not yield scores that are on the same metric. A score of 7 correct on Form A does not necessarily indicate the same level of math proficiency as a score of 7 on Form B or Form C. The reason for this is that the items of Forms A, B, and C are not necessarily of the same difficulty level. If the items of Form A are more difficult than the items of Form B, then a score of 7 correct on Form A would reflect a higher level of proficiency than the same score on Form B. The lack of a common metric for the "number correct" score across the three forms wreaks havoc for researchers attempting to measure growth across time because any change in score across Forms A, B, and C will be due not only to changes in proficiency but also to differences in the difficulty levels of the three forms. As a result, we sought an alternative methodology for creating our progress monitoring measures that would place the scores for each of the alternate forms on a common metric.

The approach we adopted was to create seven forms of the math problem solving assessment that were equated (placed on a common metric) using the measurement

Table 1
Display of items contained in each of the seven CBM forms.

Group	Item	Form 1	Form 2	Form 3	Form 4	Form 5	Form 6	Form 7
G1	1	✓			✓			✓
	2	✓			✓			✓
	3	✓			✓			✓
	4	✓			✓			✓
	5	✓			✓			✓
G2	6	✓				✓		
	7	✓				✓		
	8	✓				✓		
	9	✓				✓		
	10	✓				✓		
G3	11		✓		✓			✓
	12		✓		✓			✓
	13		✓		✓			✓
	14		✓		✓			✓
	15		✓		✓			✓
G4	16		✓				✓	
	17		✓				✓	
	18		✓				✓	
	19		✓				✓	
	20		✓				✓	
G5	21			✓		✓		
	22			✓		✓		
	23			✓		✓		
	24			✓		✓		
	25			✓		✓		
G6	26			✓			✓	
	27			✓			✓	
	28			✓			✓	
	29			✓			✓	
	30			✓			✓	

modeling framework of item response theory (Lord, 1980). The seven forms (Form 1, Form 2, …, Form 7) would be administered sequentially across time points 1 to 7. The first stage of developing the seven equated forms was to divide the 30 item pool into six 5-item groups. We refer to these six 5-item groups as G1 (items 1–5), G2 (items 6–10), G3 (items 11–15), G4 (items 16–20), G5 (items 21–25), and G6 (items 26–30). From these six 5-item groupings, we created seven 10-item forms. Form 1 contained the items from G1 and G2, Form 2 contained the items from G3 and G4, Form 3 contained the items from G5 and G6, Form 4 contained the items from G1 and G3, Form 5 contained the items from G2 and G5, Form 6 contained the items from G4 and G6, and Form 7 contained the items from G1 and G3. A tabular representation of the item composition across all seven forms is displayed in Table 1. Notice that the seven forms have overlapping items, but only two of the seven forms have identical items (Forms 4 and 7).

The common (overlapping) items across the seven forms allowed us to use a traditional common-item nonequivalent groups design to equate across the seven forms (Kolen & Brennan, 1987), which served to place the item difficulty parameters (the higher the difficulty parameter, the more difficult the item) obtained from each of the seven forms on a common metric. Because these seven forms were completed by students having different levels of math proficiency (due to maturation and any effects of the treatment), the item difficulty parameter values obtained in their non-equated form are on different metrics across the seven forms. That is, the difficulty parameter for each item will vary depending on the math proficiency of the group administered the particular form of the assessment (i.e., the same item will have a different difficulty parameter when administered at different time points), as well as the difficulty of the other items on the assessment. The goal of the equating analysis was to transform the item difficulty parameters appropriately such that all item difficulty parameter values were on the same metric. Once all item difficulties of the seven assessment forms are placed on a common metric, then the ability estimates obtained across all seven time points are on a common metric and can be compared. In our study, we used the metric of the initial item parameter estimates for Form 1 as the metric to which all other forms were equated. The equating analysis then aimed to place the difficulty parameters of Forms 2 to 7 on the same metric as Form 1.

To accomplish the equating analysis, we employed the dichotomous Rasch model (Rasch, 1980) in all item calibrations. The Rasch model is a probabilistic measurement model that specifies the probability of correct response to a test item as a function of examinee ability and the difficulty parameter of the item. Examinee ability is estimated as a function of the number of correct responses to the items on the test in addition to the difficulty of the items on the test. Details on the Rasch model are not presented here, but interested readers are referred to Andrich (1988), Bond and Fox (2001), and Fischer and Molenaar (1995) for more information.

The equating methodology employed in this study began by conducting an initial Rasch calibration of all seven forms and then recording the resulting non-equated item difficulty parameter estimates. Because Form 1 served as the metric to which all other forms were equated, the difficulty parameter estimates of Form 1 from the initial Rasch calibration were fixed to their initial values and never transformed. The remainder of the equating methodology consisted of placing the item difficulty parameter estimates of Forms 2–7 on the same metric as the items of Form 1 using an iterative linking procedure comprised of seven steps. Step 1

involved transforming the item difficulty parameter estimates of Form 4 so that they would be on the same metric as Form 1 using the common items of Form 1 and Form 4 (items 1–5 in this case). Note that Step 1 involves transforming the item difficulties of Form 4, rather than Form 2, because of common items (items 1–5) shared between Form 4 and Form 1. Step 2 involved transforming the item difficulty parameter estimates of Form 2 so that they would be on the same metric as Form 4 (which now share the metric of Form 1 via Step 1) using the common items of Form 4 and Form 2 (items 11–15 in this case). Step 3 involved transforming item difficulty parameter estimates of Form 5 so that they would be on the same metric as Form 1 using the common items of Form 1 and Form 5 (items 6–10 in this case). Step 4 involved transforming the item difficulty parameter estimates of Form 3 so that they would be on the same metric as Form 5 (which now share the metric of Form 1 via Step 3). Step 5 involved fixing the item difficulty parameter estimates of Form 6 to the equated values obtained in Step 2 (for items 16–20) and Step 4 (for items 26–30). Step 6 involved fixing the item difficulty parameter estimates of Form 6 to the equated values obtained in Step 2 (for items 16–20) and Step 4 (for items 26–30). Step 7 involved fixing the item difficulty parameter estimates of Form 7 to the equated values obtained for Form 4 in Step 1 (as the items of Form 7 are identical to those of Form 4).

In the linking design presented above, Form 4 and Form 7 contained identical items. While not ideal, the equivalence of these two forms was a result of the small number of items in the total item bank being distributed across the seven assessments and our goal of maintaining as long a period as possible between the administration of any single group of items. By the end of the sixth assessment, each item in the 30-item item bank had been administered twice (i.e., each item had occurred in two different forms). As a result, Form 7 was forced to include items that had already been administered two times, and our goal was to select the items for which the greatest time had passed since their last administration. The items of groups G1 and G3 (the same items of Form 4) best satisfied this goal, leading Form 7 to be identical to Form 4.

Having transformed and/or fixed the item parameters for Forms 2–7 so that they were all on the metric of Form 1, ability estimates were obtained using the Rasch model for each of the seven forms given across the seven time points. The resulting ability estimates could then be compared across all seven forms, even though the different forms contained items of differing difficulty. Note that these equated Rasch ability estimates, not the raw "number correct" test scores, were used in all statistical modeling of growth for students in the pilot study. The advantages of using equated scores are described in the next section.

CBM analysis example

The previous sections underscore the importance of expressing CBM scores on a common metric when performing longitudinal analyses. To reiterate, raw scores (e.g., number of correct responses) are problematic because they fail to account for the differences in item difficulties that typically arise from alternate test forms. The equating methodology that we described in the previous section eliminates this problem by linking the alternate forms to a common score metric (in this case, the scale of the baseline assessment). To illustrate the differences that can arise from using raw versus equated scores, we used the first-year pilot data from the *Solve It!* intervention to estimate a multilevel growth curve model. The measures were administered to participating students in the intervention group seven times during their math class,

specifically, prior to the intervention (baseline) and then monthly for the remainder of the school year (progress monitoring). The measures were administered three times to the participating comparison group students prior to the intervention and then at the second and seventh administrations. The internal consistency of the measures ranged from .72 to .88 for the pilot study (Montague, Enders, & Dietz, 2009).

The multilevel growth model expresses the outcome variable as a function of a temporal predictor variable that captures the passage of time. The basic linear growth model is

$$\mathrm{CBM}_{ti} = \gamma_{00} + \gamma_{10}(\mathrm{TIME}_{ti}) + u_{0i} + u_{1i}(\mathrm{TIME}_{ti}) + r_{ti} \quad (1)$$

where CBM_{ti} is the outcome score for case i at time t, $TIME_{ti}$ is the value of the temporal predictor for case i at time t (e.g., months), γ_{00} is the intercept, γ_{10} is the expected change in the outcome variable for a one-unit increment in the *TIME* variable, u_{0i} and u_{1i} are residuals that allow the intercepts and the slopes to vary across individuals, and r_{ti} is a time-specific residual. The growth model does not estimate the residuals themselves, but rather the variance of the residuals. For example, the variance of u_{0j} quantifies individual differences in math problem solving at the baseline assessment, and the variance of u_{1j} captures the degree to which changes in performance vary across individuals. Readers that are interested in additional details on the growth model can consult the Peugh (2010) manuscript in this issue.

In the current study, the *TIME* variable was centered to reflect the number of months preceding the final data collection wave. The assessments were not exactly equally spaced, so the values of *TIME* reflected this fact (*TIME* = −6.26, −5.24, −4.13, −3.28, −1.93, −.78, and 0). Coding the final assessment with a zero value facilitates the interpretation of the intercept parameter, such that γ_{00} represents the Wave 7 mean. The primary goal of the study was to determine whether participation in a novel intervention had an influence on the rate of change of the *CBM* measure. To assess the effectiveness of the intervention, a dummy code (0 = comparison group, 1 = intervention group) was added as a predictor of the intercept and slope parameters. The resulting model is

$$\begin{aligned}\mathrm{CBM}_{ti} = {} & \gamma_{00} + \gamma_{10}(\mathrm{TIME}_{ti}) + \gamma_{01}(\mathrm{INTERVENTION}_i) \\ & + \gamma_{11}(\mathrm{INTERVENTION}_i)(\mathrm{TIME}_{ti}) + u_{0i} + u_{1i}(\mathrm{TIME}_{ti}) + r_{ti},\end{aligned} \quad (2)$$

where γ_{00} and γ_{10} are the intercept (i.e., Wave 7 mean) and slope (i.e., monthly growth rate) for the comparison group, respectively, γ_{01} is the Wave 7 mean difference between the two groups, and γ_{11} is difference in growth rates between the intervention and the comparison groups. The γ_{01} and γ_{11} coefficients are of particular interest because they quantify the intervention effect (i.e., the mean differences and the end of the study and the group-by-time interaction).

To illustrate the impact that equating can have on longitudinal assessments of change, we used both the raw (i.e., number of problems correct) CBM scores and the Rasch model ability estimates to estimate the growth curve model in Eq. (2). Table 2 gives the resulting parameter estimates from both analyses. Note that the score metrics for the two analyses are different, so the point estimates from the two scaling methods are not directly comparable (raw scores reflect the number of correct problems out of 10 whereas the ability estimates

Table 2
Growth curve parameter estimates.

Parameter	Est.	*SE*	*p*
Raw scores			
Comparison Wave 7 mean (γ_{00})	6.318	.237	<.001
Comparison growth rate (γ_{10})	.338	.037	<.001
Wave 7 Mean difference (γ_{01})	.236	.289	.415
Growth rate difference (γ_{11})	−.107	.046	.019
Intercept variance (τ_{00})	3.019	.312	<.001
Slope variance (τ_{11})	N/A	N/A	N/A
Intercept/slope covariance (τ_{10})	N/A	N/A	N/A
Residual variance (σ^2)	3.543	.149	<.001
Equated scores			
Comparison Wave 7 mean (γ_{00})	.286	.174	.101
Comparison growth rate (γ_{10})	−.010	.024	.687
Wave 7 mean difference (γ_{01})	1.095	.216	<.001
Growth rate difference (γ_{11})	.116	.029	<.001
Intercept variance (τ_{00})	2.241	.258	<.001
Slope variance (τ_{11})	.006	.005	.193
Intercept/slope covariance (τ_{10})	.111	.030	<.001
Residual variance (σ^2)	1.310	.061	<.001

are on a metric similar to a *z* score). Consequently, we will focus on the substantive interpretation of the two analyses.

To begin, the raw score analysis indicated that the comparison group improved by roughly one-third of a point per month, which was a statistically significant gain, $\gamma_{10}=.338$, $p<.001$. More importantly, the group-by-time interaction was significant, indicating that the change

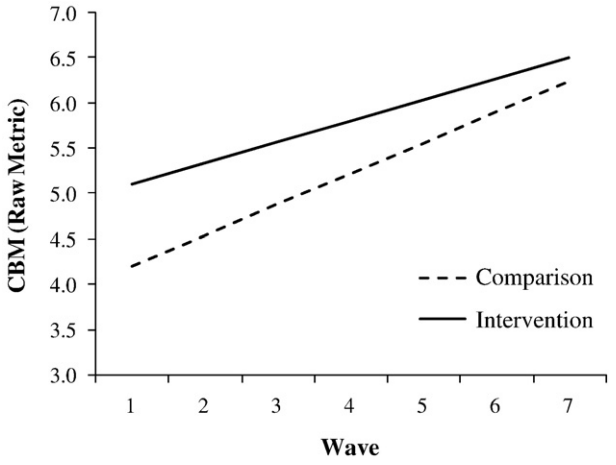

Fig. 1. Average growth curves from the analysis of raw CBM scores. The comparison growth rate (i.e., monthly change rate) was statistically significant, as was the group-by-time interaction. The intervention group change rate was less than that of the comparison group, which produced a non-significant mean difference at Wave 7.

rate for the intervention group was different than that of the comparison group, $\gamma_{11}=-.107$, $p<.019$. However, the negative sign of this coefficient suggests that the intervention group actually improved at a slower rate than the comparison group; the intervention group growth rate was slightly less than a quarter of a point per month, $\gamma_{10}+\gamma_{11}=.338-.107=.231$. To further illustrate these results, we used the regression coefficients to compute the simple slopes (i.e., model-predicted means). Fig. 1 shows the average growth curve for the two conditions. The significant group-by-time interaction is evidenced by the slope difference in the two trajectories, and the vertical separation of the growth curves at Wave 7 represents the mean difference, which was not significant, $\gamma_{01}=-.236$, $p=.415$. From a substantive standpoint, Fig. 1 suggests that the comparison group effectively "caught up" with the intervention group by the end of the study, such that there was no material difference between the groups. Not surprisingly, this finding is inconsistent with expectations.

The bottom portion of Table 2 gives the growth model parameter estimates from the equated CBM scores. The analysis of the Rasch ability estimates produced a very different substantive conclusion. Specifically, the average monthly change rate for the comparison group was non-significant, $\gamma_{10}=-.010$, $p=.687$ (the value of the coefficient effectively represents monthly change on the z score metric). Consistent with the raw score analysis, the group-by-time interaction was significant, indicating that the change rate for the intervention group was different than that of the comparison group, $\gamma_{11}=.116$, $p<.001$. However, the sign of the coefficient was positive in this analysis, meaning that the intervention group showed greater improvement over time relative to the comparison group (the intervention group growth rate was $\gamma_{10}+\gamma_{11}=-.010+.116=.106$). Fig. 2 shows the average growth curve for the two conditions. The significant group-by-time interaction is evidenced by the slope difference in the two trajectories, and the vertical separation of the growth curves at Wave 7 represents the mean difference, which was statistically significant,

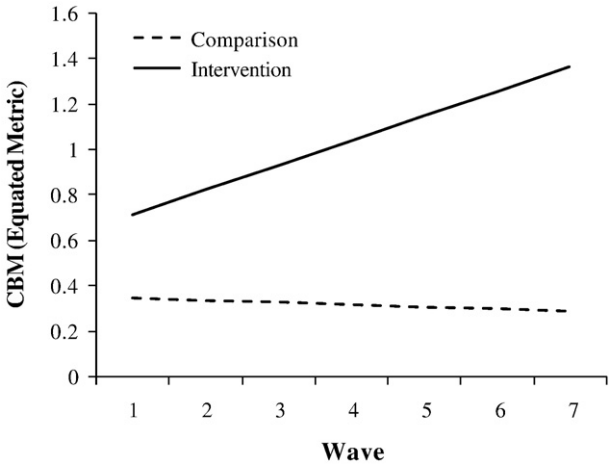

Fig. 2. Average growth curves from the analysis of equated CBM scores. The comparison growth rate (i.e., monthly change rate) was non-significant, but the group-by-time interaction was significant, indicating that the intervention group improved relative to the comparison group. The growth rate difference produced a significant mean difference between the groups at Wave 7.

$\gamma_{01}=1.095$, $p<.001$. From a substantive standpoint, Fig. 2 is consistent with the expectation that the intervention group improved relative to the comparison group.

Discussion

There are several ways to interpret the differences between the results using the raw test scores (number of correct problems out of ten) and Rasch measures of math ability. One explanation resides in two distinctions between the properties of raw test scores and the Rasch measures of math ability. The first distinction concerns the relationship between the raw scores and the underlying ability. The raw score is nonlinearly related to the underlying ability metric, such that an increase of one unit in the raw score (i.e., an increase of one more item correct out of ten) corresponds to different increases in actual ability depending on the raw score. That is, a difference in math ability between raw scores of 1 and 2 is different than the difference in math ability between raw scores 5 and 6 (in general, one would expect the difference between raw scores of 1 and 2 to reflect a larger difference in math ability than that of the difference in raw scores of 5 and 6). As a result of this property, it is possible to observe negligible gains in raw score despite there being substantial gains in underlying math ability (although this will depend on the initial level of ability in relation to the difficulty of the test). In particular, if the math ability distributions differ for the treatment and comparison groups (as they would if there was any treatment effect over time), then it is possible for the raw score changes across time to be similar for the two groups yet to have between-group differences in underlying math ability being estimated by the Rasch analysis.

A second distinction between the raw score analysis and the analyses based on the Rasch estimates of math ability concerns the difficulty of the assessments used across the seven time points. The raw scores obtained at each of the seven time points are not linked to a common metric, so the raw scores at different time points depend on the difficulty of the particular items administered at each assessment. That is, a particular raw score (e.g., a score of 7 correct out of 10) does not reflect the same level of math ability across the seven time points. As a result, a gain in raw test scores from one time point to the next may not reflect a gain in math ability if the difficulty of the items decreased across the time points. That is, a gain on the raw score metric might simply reflect the fact that the items from the later assessments were less difficult than the items from the earlier assessments. In contrast, the ability estimates from the Rasch analysis were on a common metric at every assessment, so the growth model parameter estimates did not reflect idiosyncratic differences in the psychometric characteristics of the prompts. In the growth analysis based on the Rasch ability estimates (see Fig. 2), we observed that the mean ability remained relatively constant for the comparison group across the seven time points, which was expected because the mean ability of the comparison group was not expected to increase substantially across time. This contrasts with the results obtained using the raw scores (Fig. 1) where the scores for the comparison group increased over time (even more so than the intervention group).

It is important to point out that the pattern of results that we illustrated in our growth model analyses is specific to our particular study. The accuracy of a raw score analysis will depend on the degree of nonlinearity between the raw scores and the latent ability metric as well as on the pattern of item difficulties across time. Both of these factors are likely to vary

from study to study. Consequently, it is reasonable to expect even larger discrepancies between the raw score and Rasch metrics in some situations and smaller discrepancies in other cases. However, given that the Rasch analysis yields estimates of math ability that are linearly related to the underlying latent math ability continuum and are on a common metric across time, a strong argument can be made that the longitudinal analyses based on the Rasch ability estimates provide a more accurate assessment of the true intervention effect.

In conclusion, we attempted to present an argument in favor of reforming the present approach to CBM. As part of this reform, we advocate the use of equated scores rather than raw scores for CBM that more accurately reflect a student's academic performance across time. Certainly this is essential for research purposes, but we also underscore the importance of this approach for educational decision-making. CBM (i.e., progress monitoring) in general education classrooms is rapidly becoming the norm as the basis for making crucial educational decisions about individual children. The validity of these decisions made by educators for students across their school years depends on accurate and technically sound data. CBM should be held to high standards and accurately reflect the performance of students that ultimately will lead to more informed decision-making.

Acknowledgments

This research is supported by grant no. R324A070206 from the Institute for Education Sciences (IES), U.S. Department of Education. Opinions expressed herein are those of the authors and do not represent the position of the U.S. Department of Education. The authors are grateful to the students as well as the teachers, administrators, and other school personnel in the Miami-Dade County Public Schools for their cooperation and support.

References

Andrich, D. (1988). Rasch models for measurement Newbury Park: Sage.

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.

Bond, T. G., & Fox, C. M. (2001). Applying the Rasch model: Fundamental measurement in the human sciences New Jersey: Lawrence Erlbaum.

Busch, T. W., & Reschly, A. L. (2007). Progress monitoring in reading. *Assessment for Effective Intervention*, *32*, 223−230.

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.

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

Deno, S. L. (2003). Developments on curriculum-based measurement. *The Journal of Special Education*, *37*, 184−192.

Deno, S. L., & Fuchs, L. S. (1987). Developing curriculum-based measurement systems for data-based special education problem solving. *Focus on Exceptional Children*, *19*, 1−15.

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

Deno, S. L., & Mirkin, P. K. (1977). Data-based program modification: A manual Reston, VA: Council for Exceptional Children.

Fischer, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch models: Foundations, recent developments, and applications* New York: Springer-Verlag.

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *The Journal of Special Education*, *41*, 121−139.

Foegen, A., & Deno, S. L. (2001). Identifying growth indicators for low-achieving students in middle school mathematics. *The Journal of Special Education*, *35*, 4−16.

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

Fuchs, L. S., & Deno, S. L. (1992). Effects of curriculum within curriculum-based measurement. *Exceptional Children*, *58*, 232−242.

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

Fuchs, L. S., Fuchs, D., & Courey, S. J. (2005). Curriculum-based measurement of mathematics competence: From computation to concepts and applications to real-life problem solving. *Assessment for Effective Intervention*, *30*, 33−46.

Fuchs, L. S., Fuchs, D., & Hamelett, C. L. (1989). Effects of alternative goal structures within curriculum-based measurement. *Exceptional Children*, *55*, 429−438.

Fuchs, L. S., Fuchs, D., Hamelett, C. L., & Allinder, R. M. (1989). The reliability and validity of skills analysis within curriculum-based measurement. *Diagnostique*, *14*, 203−221.

Fuchs, L. S., Fuchs, D., Hamelett, C. L., & Stecker, P. M. (1990). The role of skills analysis in curriculum-based measurement in math. *School Psychology Review*, *19*, 6−22.

Fuchs, L. S., Fuchs, D., Hamelett, C. L., & Stecker, P. M. (1991). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations. *American Educational Research*, *28*, 617−641.

Fuchs, L. S., Fuchs, D., Hamelett, C. L., Thompson, A., Roberts, P. H., Kubek, P., et al. (1994). Technical features of a mathematics concepts and applications curriculum-based measurement system. *Diagnostique*, *19*, 23−49.

Helwig, R., & Tindal, G. (2002). Using general outcome measures in mathematics to measure adequate yearly progress as mandated by Title I. *Assessment for Effective Intervention*, *28*, 9−18.

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

Kolen, M. J., & Brennan, R. L. (1987). Linear equating models for the common item nonequivalent populations design. *Applied Psychological Measurement*, *11*, 263−277.

Lord, F. M. (1980). Applications of item response theory to practical testing problems NJ: Lawrence Erlbaum.

Madelaine, A., & Wheldall, K. (1999). Curriculum-based measurement of reading: A critical review. *International Journal of Disability, Development, and Education*, *46*, 71−85.

Montague, M. (2003). Solve it! A mathematical problem-solving instructional program Reston, VA: Exceptional Innovations.

Montague, M., Enders, C., & Dietz, S. (2009). The effects of Solve It! on middle school students' math problem solving and math self-efficacy. Manuscript submitted for publication.

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*.

Rasch, G. (1980). Probabilitic models for some intelligence and attainment tests (Expanded ed.) Chicago: University of Chicago Press.

Shapiro, E. S., Edwards, L., & Zigmond, N. (2005). Progress monitoring of mathematics among students with learning disabilities. *Assessment for Effective Intervention*, *30*, 15−32.

Speece, D. L., Case, L. P., & Molloy, D. E. (2003). The validity of a response-to-instruction paradigm to identify reading disabilities: A longitudinal analysis of individual differences and contextual factors. *School Psychology Review*, *32*, 557−582.

Wallace, T., Espin, C. A., McMaster, K., Deno, S. L., & Foegen, A. (2007). CBM progress monitoring within a standards-based system. *The Journal of Special Education*, *41*, 66−67.