Abstract Title Page

Title: Curriculum-based Measurement of Math Problem Solving: A Methodology and Rationale for Establishing Equivalence of Scores

Author(s): Marjorie Montague
Randall Penfield
Craig Enders
Jia Huang

Abstract Body

Background/context:
Curriculum-based measurement (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). 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 making critical educational decisions for students (e.g., special education placement). Our CBM-M research focuses on the development of seven alternate forms that produce statistically equivalent scores to measure growth in math problem solving.

Purpose / objective / research question / focus of study:
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 using IRT methodology. The research context for the development of these curriculum-based measures is a federally funded Goal 3 IES 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 middle school 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. The math word problems for the alternate forms of the measures were selected from the *Solve It!* manual (Montague, 2003). Each measure consisted of 10 one-, two-, and three-step textbook-style word 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.

Setting:

The study was conducted in the Miami-Dade County Public Schools, the 4th largest school district in the nation, which serves about 370,000 students (10 % white, 29 % African-American, 59 % Hispanic, and 2 % other).

Population / Participants / Subjects:

Participants were 312 middle school students (average achieving, *n* = 100; low achieving, *n* = 159; students with learning disabilities, *n* = 53).

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 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. (Please insert Table 1 here.) 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, 1995), 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. 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. Step1 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, (i.e., prior to the intervention and 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). Multilevel growth modeling was used to capture progress over time. 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 Equation 2. Table 2 gives the resulting parameter estimates from both analyses. (Please insert Table 2 here.) 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 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, *10* = .338, *p* < .001. More importantly, the group-by-time interaction was significant, indicating that the change rate for the intervention group was different than that of the comparison group, *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, *10* + *11* = .338 - .107 = .231. To further illustrate these results, we used the regression coefficients to compute the simple slopes (i.e., model-predicted means). Figure 1 shows the average growth curve for the two conditions. (Please insert Figure 1 here.) 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, *01* = -.236, *p* = .415. From a substantive standpoint, Figure 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, *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, *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 *10* + *11* = -.010 + .116 = .106). Figure 2 shows the average growth curve for the two conditions. (Please insert Figure 2 here.) 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, *01* = 1.095, *p* < .001. From a substantive standpoint, Figure 2 is consistent with the expectation that the intervention group improved relative to the comparison group.

Interpretations/Conclusions: In conclusion, we argue that the present approach to CBM needs reform. 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.

Appendices

Appendix A. References:

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.

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., & 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.

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.

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

Appendix B. Tables and Figures

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			✓			✓	

Table 2

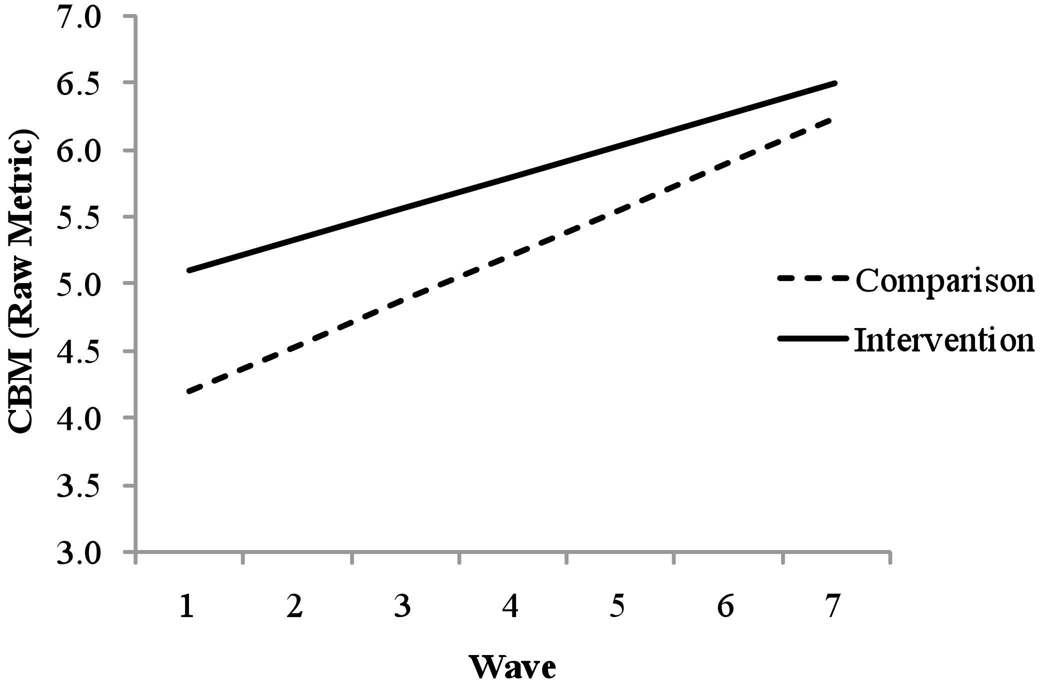
Growth Curve Parameter Estimates

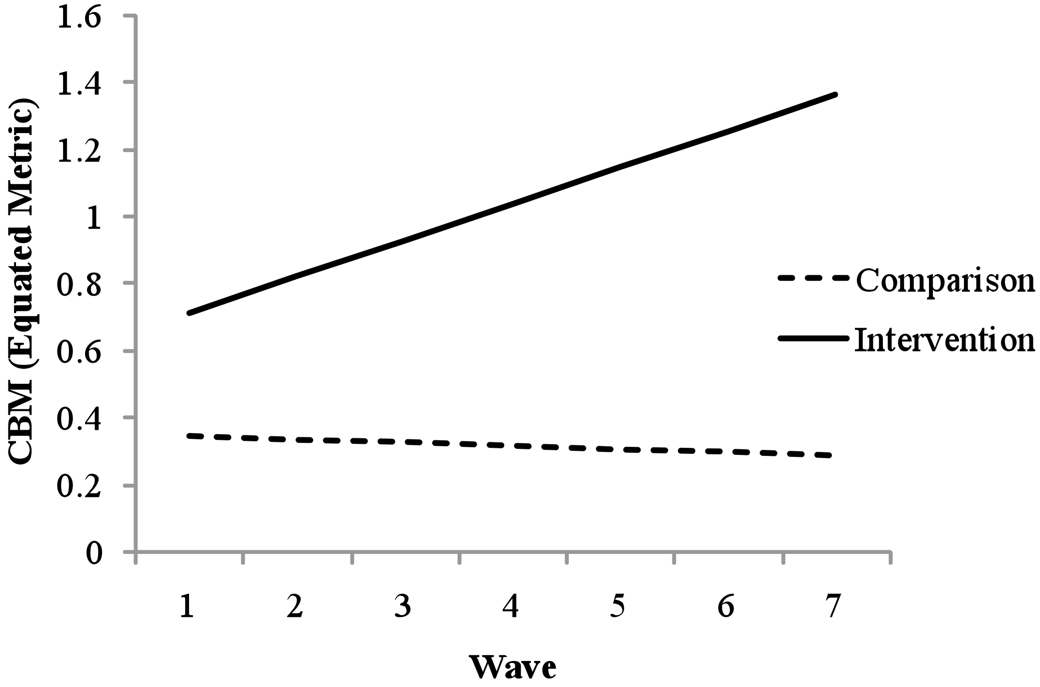
Parameter	Est.	*SE*	*P*
Raw Scores			
Comparison Wave 7 Mean (⍰*00*)	6.318	0.237	< .001
Comparison Growth Rate (⍰*10*)	0.338	0.037	< .001
Wave 7 Mean Difference (⍰*01*)	0.236	0.289	.415
Growth Rate Difference (⍰*11*)	-0.107	0.046	.019
Intercept Variance (⍰*00*)	3.019	0.312	< .001
Slope Variance (⍰*11*)	N/A	N/A	N/A
Intercept/Slope Covariance (⍰*10*)	N/A	N/A	N/A
Residual Variance (⍰*2*)	3.543	0.149	< .001
Equated Scores			
Comparison Wave 7 Mean (⍰*00*)	0.286	0.174	.101
Comparison Growth Rate (⍰*10*)	-0.010	0.024	.687
Wave 7 Mean Difference (⍰*01*)	1.095	0.216	< .001
Growth Rate Difference (⍰*11*)	0.116	0.029	< .001
Intercept Variance (⍰*00*)	2.241	0.258	< .001
Slope Variance (⍰*11*)	0.006	0.005	.193
Intercept/Slope Covariance (⍰*10*)	0.111	0.030	< .001
Residual Variance (⍰*2*)	1.310	0.061	< .001

Figure Captions

Figure 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.

Figure 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.


7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
CBM (Raw Metric)
1
2
3
4
5
6
7
Wave
Comparison
Intervention


1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
CBM (Equated Metric)
1
2
3
4
5
6
7
Wave
Comparison
Intervention