

RUNNING HEAD: Within-Year Math Growth

Within-year Growth in Math: Implications for Progress-Monitoring Using RTI

P. Shawn Irvin, Daniel Anderson, Jessica L. Saven, Julie Alonzo, and Gerald Tindal

University of Oregon

Acknowledgements:

Note: Funds for the data set used to generate this report came from a federal grant awarded to the UO from the Institute of Education Sciences, U.S. Department of Education: Developing Middle School Mathematics Progress Monitoring Measures (R324A100026 funded from June 2010 - June 2014).

### Abstract

The current study is framed in the context of monitoring student progress that is increasingly couched within a Response to Intervention (RTI) approach to teaching and learning. Using a three-level hierarchical linear model (HLM), we examine initial status and growth for fourth grade students who received short mathematics progress-monitoring assessments during the 2011-2012 school year. We examine student growth by examining extant data that reflect teacher practices in the field. Two HLM analyses are presented. The first is conducted with a national sample of convenience, and is designed to investigate and characterize the amount and type of within-year growth in math. The second analysis, more exploratory in nature, is conducted with a further restricted sample, and is designed to examine the relationship between three important indicators of student literacy and initial status and within-year growth in math. Findings indicate significant within-year growth in math, though the nature of the growth trajectory is unclear. Additionally, while student growth is statistically observable, given the amount of growth relative to the scale of the progress-monitoring probe used as the outcome in our analyses, the usefulness of raw score reporting as a guide for instructional decision-making is also discussed. Further research on within-year growth in math is suggested, especially as it relates to the influence of various instructional practices on such growth, an important area of interest not addressed in this study.

## **Within-year Growth in Math: Implications for Progress-Monitoring Using RTI**

Public schools face challenges to improve student learning and achievement. Educators implementing school-wide improvement efforts such as Response to Intervention (RTI; D. Fuchs & Fuchs, 2001) attempt to meet the diverse learning needs of students performing below expectations. The enactment of this goal, in part through ongoing progress monitoring aimed at characterizing within-year student progress (growth), is the focus of this study.

RTI can be viewed as a grassroots school-wide improvement effort based on results of instructional interventions as measured by interim and formative assessments (Black & Wiliam, 1998). In RTI, students are classified as "at-risk" of not meeting grade-level expectations through interim screening assessments, typically administered seasonally during the academic year. Students performing below a district- or school-designated level on interim screeners are provided an instructional intervention designed to improve achievement and are administered frequent formative progress-monitoring probes to track the effect of the intervention and ensure adequate gains (D. Fuchs, Mock, Morgan, & Young, 2003). RTI is intended to be dynamic, with students receiving individualized intervention based on their learning needs. The RTI process ideally results in students' academic deficits being identified early so that targeted interventions can be provided and achievement raised over time.

In contrast to longer statewide summative exams, formative assessments are intended to provide teachers with quick information to guide instructional decisions, rather than to evaluate the long-term results of instruction and content coverage (Black & William, 1998). Formative assessments must be brief to administer, with multiple equivalently difficult test forms available in each grade, so that teachers can efficiently evaluate students' within-year growth.

Despite the necessarily brief administration time, formative measures must also maintain adequate psychometric characteristics. Within RTI, teachers base instructional decisions on the results of formative measures. Assessments with weak technical characteristics may misguide teachers' instructional practices and threaten the validity of RTI. For instance, measures with weak psychometric characteristics may lead to some students receiving an intervention who are not necessarily in need (false positives), while others are withheld intervention despite being in need (false negatives). Regardless of these challenges, formative assessments are key for characterizing student growth within the context of RTI and guiding instructional decision-making (D. Fuchs, et al., 2003). Thus, the capacity of such assessments to adequately capture the within-year growth of students performing below expectations is essential.

In this paper, we evaluate students' within-year growth in grade 4 in a single mathematical domain – *Number and Operations*. We present two analyses using hierarchical linear modeling (Raudenbush & Bryk, 2002). First, we explore the extent to which students' fall math benchmark score predicts the intercepts (starting point) and slopes (change in performance over time) for students who were progress monitored. Second, we restrict the analysis to a subsample of students who had valid scores on three fall reading benchmark screeners assessing their oral reading fluency, reading comprehension, and vocabulary. We then explore the extent to which the reading measures predict the intercepts and slopes of students who were progress monitored in *Number and Operations*. The following research questions guided our analyses:

1. What are the typical characteristics of student growth observed with a short 16-item *Number and Operations* mathematics progress monitoring assessment in Grade 4?  
(Analysis 1)

2. Does performance on the fall math benchmark predict students' progress monitoring intercepts and/or slopes on *Number and Operations* mathematics progress monitoring assessments? (Analysis 1)
3. Do students' fall reading scores from the Grade 4 benchmark reading battery predict their progress monitoring intercepts and/or slopes on *Number and Operations* mathematics progress monitoring assessments? (Analysis 2)

The purpose of the first analysis was to explore, generally, how students progressed during the year and how the fall mathematics benchmark assessment predicted their growth. The purpose of the second analysis was more exploratory. All math items used in the study were administered online and developed with adherence to principles of Universal Design for Assessment (Thompson, Johnstone, & Thurlow, 2002). Included in the delivery of items was the option for students to have item stems and response options read aloud. Because of these design features, we would theoretically expect the relation between reading and math to be low. The purpose of the second analysis was to test the relation between performance on these two content-area assessments empirically. Below, we synthesize current research related to mathematics growth, as well as universal design and the relation between reading and math.

### **Within-Year Math Growth**

Previous research has highlighted the limited knowledge of typical within-year growth in math. For example, in a review of research on mathematics progress measures, Foegen, Jiban and Deno (2007) stated “the pool of measures about which we have some evidence of their use as progress indicators is substantially smaller than is the pool of measures with evidence for their use as static indicators” (pp. 137-138). Indeed, of the 32 studies the authors identified in their review, only 9 explored any characteristic of the slope over time. More recent research by

Lembke, Hampton and Beyers (2012) suggests that the field has shifted towards a greater focus on growth. The authors identified 27 studies that used some means of calculating a weekly growth rate. Yet, all but three of these studies were at the kindergarten or first grade level, with the vast majority focusing on fluency-based indicators (e.g., the number of digits correct; see L. Fuchs, Fuchs, & Courey, 2005).

As students progress into the upper elementary and middle school grades, the mathematical concepts become more complex and less amenable to fluency-based indicators. Test construction for such purposes must then be reconceptualized in terms of the depth and breadth of mathematics constructs targeted, while maintaining valid test development practices such as Universal Design for Assessment. The inherent necessity to reconceptualize test construction for grades beyond early elementary generally leads to progress being more difficult to measure, perhaps one reason why so few studies have investigated within-year math growth beyond early elementary and/or fluency-based indicators (Calhoon, 2008; Foegen & Deno, 2001; Foegen, et al., 2007; Lembke, et al., 2012). Yet, within an RTI framework the within-year growth of students being progress monitored is critical. Further research is needed to identify test construction practices that lead to measures sensitive to detecting growth.

### **Universal Design**

Universal Design for Assessment (UDA) is a test development process that attempts to maximize the accessibility of test items to the widest range of respondents possible. UDA originated in the field of architecture as a means of designing structures to be accessible to persons with disabilities (Thompson, et al., 2002). By designing buildings with the potential users in mind from the beginning, construction can proceed in a way that makes the structure

accessible to all individuals without sacrificing visual aesthetics (e.g., wheelchair ramps). UDA adopted these principles of designing for accessibility for use in assessment development.

Thompson et al. (2002) outlined seven UDA test development principles. Two of these principles directly relate to the text students are asked to read: (a) maximum readability and comprehensibility and (b) maximum legibility. When, as in the current study, the construct being measured is *not* reading, other principles relate to excluding text, including defining constructs precisely. In other words, math items should be as free of text as possible (to ensure the construct being measured is math and not reading), and should target only a single math skill.

In the current study, all items were developed adhering to the UDA principles outlined above. Further efforts were made to ensure that reading was not a significant factor in how students performed on the mathematics items, including the availability of read-aloud options for all item stems and response options containing written words. Figure 1 displays a sample item used in this study. Note the large amount of white space, reduced text, simple line art graphic, and read-aloud buttons. The item interface also contains UDA elements. For example, students can select a response option by clicking anywhere on the rectangular box representing the option. The response option then highlights blue (as shown in Figure 1) and the student can proceed to the next option by selecting the large “Next” button (again, clicking anywhere around the button). Given these design elements and the overall reduction in text/reading required of students, we were interested in exploring the extent to which reading measures would predict students’ performance. Previous research has suggested the relation between reading and math is quite strong (e.g., Stevens, 1995). We thus expected the relation to be present, but hypothesized that the magnitude of the relation would be suppressed.

In what follows, we present our methods for investigating student growth on a math progress monitoring measure during the 2011-2012 school. Results from this portion of the study serve to broaden the research base around within-year growth in math to include findings for later elementary students who are instructed in more complex mathematical concepts (beyond that of fluency-based numeracy skills).

### **Methods**

In what follows, we organize our methods by the data cleaning steps taken and resulting analytic samples, the measures used in the study, and the two HLM analyses conducted at the fourth grade level.

#### **Data Cleaning and Analytic Samples**

The full extant database contained 29,728 students from 563 schools across 122 districts and 27 states. For this study, we were primarily interested in the growth made by students who were progress monitored using the easyCBM *Number and Operations* mathematics measure. However, our interest in exploring the relation between initial performance on the benchmark measures and within-year growth on the progress monitoring measures led us to restrict the sample to only students with a valid benchmark score who were progress-monitored. This step reduced the total sample size to 10,018 – or approximately 34% of the total sample. This number aligns well with expectations that students performing below the normative 25th percentile be progress-monitored Fuchs and Fuchs (2006). Restricting the sample to only students who had at least two progress monitoring scores in the area of *Number and Operations* and a valid benchmark score resulted in our final analytic sample for Analysis 1 of 2,189 students representing 105 schools. For Analysis 2, the sample was further reduced to include only students who had valid passage reading fluency, comprehension, and vocabulary scores on the



fall reading benchmark assessments, resulting in a final sample of 474 students in 42 schools for this second analysis.

## Measures

Two mathematics and three reading measures were used in this study. For mathematics, we used the fourth-grade easyCBM *Number and Operations* progress monitoring probes and the fall math interim-benchmark screener. In reading, we used three easyCBM interim-benchmark screeners: passage reading fluency, reading comprehension, and vocabulary. Below, we summarize the technical adequacy of each measure used.

**Number and operations.** All easyCBM math progress-monitoring forms were written to align with each of the National Council of Teachers of Mathematics (NCTM) Focal Point Standards (Alonzo, Lai, & Tindal, 2009) and designed to be of equivalent difficulty with a Rasch model (Alonzo, Anderson, & Tindal, 2009). For grade 4, the NCTM standards include *Number and Operations*, *Measurement*, and *Number and Operations and Algebra*. These easyCBM probes were selected because they were designed to measure higher-order, more complex mathematical skills, which adds to a research base primarily focused on monitoring early developmental areas (see L. Fuchs, et al., 2005; Lembke, et al., 2012). Ten alternate progress monitoring test forms were available, all of which were designed to be of equivalent difficulty. Students' scores from these alternate forms were included as the outcome for both Analysis 1 and 2. The measures included 16 multiple-choice items on each form addressing the objectives outlined by the fourth grade *Number and Operations* focal point standard. The technical adequacy of the measures has been reported by Nese et al. (2010) and is further detailed below.

**Mathematics interim-benchmark.** The easyCBM math interim-benchmark assessment was designed for computer-based seasonal administration (fall, winter, and spring). The

assessment is comprised of 45 multiple-choice items. The benchmark tests are designed to screen students for academic risk of failing to meet grade-level expectations, and thus offer educators evidence to implement instructional interventions along with targeted progress monitoring within an RTI framework. In other words, within RTI, results from the interim-benchmark measures ideally work in conjunction with progress monitoring measures such as the *Number and Operations* probes used in this study.

Nese et al. (2010) reported internal reliability of the easyCBM math measures using Cronbach's alpha and split-half estimates. Alpha coefficients for the full sample of students included in the study were above .80, while split-half estimates were above .70. Alpha coefficients and split-half estimates remained consistent and high across specific student populations, including students who received special education services, English language learners, and race/ethnicity sub-samples. Criterion related validity for the math benchmarks with respect to the Oregon and Washington state tests has been reported, with Pearson correlation coefficients above .70 across the Grade 4 benchmarks, explaining 65% and 69% in the Oregon and Washington state tests, respectively (Anderson, Alonzo, & Tindal, 2010a, 2010b). Scores from the 2011-2012 fall math benchmark were included as predictors of student math growth in Analysis 1.

**Passage reading fluency.** The easyCBM passage reading fluency (PRF) measure is an individually-administered test of students' ability to accurately read narrative text (Alonzo & Tindal, 2007). Passages are scored based on the number of correctly read words per minute, with self-corrections counted as correct, and hesitations of longer than 3 seconds scored incorrect (with the student prompted to continue reading). For fourth grade, all passages contained

approximately 250 words of narrative text. Students' fall benchmark PRF scores were included as predictors of student math growth for Analysis 2.

**Reading comprehension.** The easyCBM multiple choice reading comprehension (MCRC) test contains 20 multiple-choice items assessing students' literal, inferential, and evaluative comprehension of narrative texts approximately 1,500 words long. The measures were designed for computer or paper-pencil group administration. Of the 20 multiple-choice items, 14 targeted either literal or inferential skills, while the remaining 6 targeted evaluative comprehension (Alonzo, Liu, & Tindal, 2007).

Saéz et al. (2010) reported reliability for the benchmark reading comprehension measures. Saez and colleagues found moderately strong evidence of reliability in Grade 4, with Cronbach's alpha ranging from .73 to .78 across all students in the study for the seasonal battery of assessments. Item correlations ranged from .60-.79 for ethnic subgroups across the three seasonal benchmarks. The reliability among English Language Learners was less consistent across grades and time points, ranging from .35-.76. Students' comprehension scores from the 2011-2012 fall benchmark were included as predictors of student math growth in Analysis 2.

**Vocabulary.** The easyCBM vocabulary (VOC) measures were developed to assess students' ability to understand the meaning of context-embedded grade-level vocabulary (Alonzo, Anderson, Park, & Tindal, 2012). The measure is group administered via computer or paper-pencil. Students silently read a single sentence containing a target vocabulary term, along with contextual information that implies the term's meaning. Directly below the context sentence, an item prompt asks students to provide the meaning of the targeted vocabulary term. Students then choose from three response options: one correct response and two plausible but incorrect distractors. The target term is bolded in both the context sentence and the item prompt

for clarification. Alonzo, Anderson, Park and Tindal (2012) reported item level fit statistics based on Rasch analysis and distractor analysis results for all items in the easyCBM vocabulary item bank, using these results to construct the alternate forms comparable forms. Students' vocabulary scores from the fall benchmark were included as predictors of student math growth in Analysis 2.

### Analyses

Two HLM analyses were conducted to explore the research questions. For both models, we began by first fitting an unconditional growth model. Visual inspection of individual students' growth slopes suggested a curvilinear trend may have fit the data better than a simple linear function. Our full a priori unconditional growth model for both analyses was thus defined as

$$\left\{ \begin{array}{l} PMScore_{tij} = \pi_{0ij} + \pi_{1ij}(time) + \pi_{2ij}(time^2) + e_{tij} \\ \pi_{0ij} = \beta_{00j} + r_{0ij} \\ \pi_{1ij} = \beta_{10j} + r_{1ij} \\ \pi_{2ij} = \beta_{20j} + r_{2ij} \\ \beta_{00j} = \gamma_{000} + u_{00j} \\ \beta_{10j} = \gamma_{100} + u_{10j} \\ \beta_{20j} = \gamma_{200} + u_{20j} \end{array} \right. \quad (1)$$

Where  $MScore_{tij}$  represents students' *Number and Operations* progress-monitoring score at time  $t$  for student  $i$  in school  $j$ . *Time* represents the fractional months (coded to represent the specific day the test was administered) occurring between test administrations, which was entered to represent a linear trend ( $\pi_{1ij}$ ) and squared to represent a curvilinear trend ( $\pi_{2ij}$ ). *A priori*, all parameters were assumed to vary between students ( $r_{0ij}$  to  $r_{2ij}$ ) and schools ( $u_{00j}$  to  $u_{20j}$ ).

The time variable was entered individually centered. That is, for each student, the first progress-monitoring occasion was coded 0, regardless of when during the academic school year the administration occurred. All subsequent time points represented the fractional months occurring between the first assessment occasion and the  $t$  assessment occasion. The model intercept thus represented the average progress monitoring score for students at the time progress monitoring began. Seasonal control variables, representing whether the student began progress monitoring the winter or spring, as opposed to the fall, were then entered as student-level predictors to evaluate whether students' intercepts and/or slopes were significantly related to the time progress monitoring began.

Our full *a priori* conditional models for each analysis are displayed in Appendix A. For Analysis 1, the easyCBM math fall benchmark was entered grand-mean centered as a student-level predictor of students' intercept and both linear and quadratic slopes, along with dummy vectors representing the season in which the student began progress monitoring (fall reference group). *A priori*, all parameters were assumed to vary randomly between schools. The primary purpose of Analysis 1 was to explore the within-year growth of students who were progress-monitored on a brief (16 item) assessment addressing objectives from the *Number and Operations* NCTM focal point standard.

For Analysis 2, the easyCBM fall math benchmark was replaced by the easyCBM fall reading benchmark tests of passage reading fluency (PRF), reading comprehension (MCRC), and vocabulary (VOC). All predictors were entered simultaneously and were grand-mean centered. All parameters were, again, assumed to vary randomly between schools. The primary purpose of Analysis 2 was to evaluate the extent to which reading performance predicted students' math

intercept and/or slope, which we hypothesized would be minimal given the UDA principles adhered to during item development.

It is important to note that our *a priori* conditional models were *not* our observed final models. Rather, for each model we began by testing the fit between a linear and curvilinear model to arrive at a baseline unconditional growth model. We then built toward our theoretical full conditional models while generally dropping non-significant effects, though theory and parsimony were also considered. All analyses were run with the HLM 7 software (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011) with full information maximum-likelihood estimation. Following our final model, residuals at all levels were investigated for adherence to the underlying assumptions of HLM (Raudenbush & Bryk, 2002). These investigations revealed no violations of the assumptions.

## Results

In what follows, model-building decisions and associated results are presented for the two HLM analyses conducted in this study.

### Analysis 1

The model that included students' fall math benchmark scores as a predictor of student progress monitoring growth was fit for a sample of students who had valid scores on both the fall math benchmark and were progress monitored using the *Number and Operations* measure.

The results of the unconditional curvilinear model are displayed as *Model 1* in Table 1. The chi-square deviance test suggested the quadratic model fit the data significantly better than the linear growth model,  $\chi^2(7) = 47.35, p < .001$ . Results suggested that students began, on average, responding correctly to 10.55 out of the 16 possible questions on the *Numbers and Operations* progress monitoring measure. This intercept value varied significantly between

students,  $\chi^2(1310) = 3359.20$ ,  $SD = 2.00$ ,  $p < .001$ , and between schools,  $\chi^2(93) = 583.49$ ,  $SD = 1.13$ ,  $p < .001$ . On average, students grew at a linear rate of .27 points per month. The linear slope did not vary significantly between students, though this value did significantly vary at the school level,  $\chi^2(93) = 192.12$ ,  $SD = .30$ ,  $p < .001$ . While the quadratic growth parameter varied significantly between schools, it was not a significant predictor of students' math scores, and did not vary significantly between students. The parameter was thus removed from future models for parsimony. The ICC for the unconditional growth model suggested that 47.64% of students' intercept variance depended on students, while 15.37% depended on schools. Additionally, 4.56% of students' linear slope variance depended on students, while 2.72% depended on schools.

Dummy coded vectors representing whether the student began progress monitoring in the winter or spring, as opposed to the fall, were then entered into the model as seasonal controls. All variables were entered as student-level predictors of their intercept and both slopes, and allowed to vary randomly between schools. The random effects for the seasonal control variables did not vary significantly between schools, and were thus fixed. The resulting model with fixed effects for the seasonal controls fit the data significantly better than the model with the random effects,  $\chi^2(20) = 107.78$ ,  $p < .001$ . The results of this model are displayed in Table 1 as *Model 2*. The winter and spring seasonal start variables were significant predictors of students' intercepts. On average, students who began progress monitoring in the winter scored 0.66 points higher than students who began in the fall. Similarly, students who began progress monitoring in the spring scored, on average, 2.36 points higher than students who began in the fall. However, the fixed effects of the seasonal control parameters were nonsignificant predictors of students' growth, and thus, were removed from the model. Overall, the model including the seasonal control

parameters accounted for 27.21% of the student-level variance beyond the unconditional growth model.

Students' scores on the fall math benchmark were then entered grand-mean centered into the model as predictors of students' intercept and linear slope, and allowed to vary randomly between schools. Overall, the final model fit significantly better than the seasonal control model,  $\chi^2(7) = 1881.75, p < .001$ , and suggested that students began, on average, responding correctly to 10.68 math questions. This value varied significantly between students,  $\chi^2(1923) = 2792.22, SD = 0.99, p < .001$ , and between schools,  $\chi^2(93) = 462.38, SD = .60, p < .001$ . Students progressed, on average, .20 points correct per month. This value varied randomly between students,  $\chi^2(2018) = 2281.10, SD = 0.14, p < .001$ , and between schools,  $\chi^2(93) = 194.88, SD = 0.05, p < .001$ .

Both seasonal control variables persisted as significant predictors of students' intercept. On average, students who began progress monitoring in the winter scored .56 points higher compared to students who began in the fall, while students who began being progress monitored in the spring scored, on average, 1.79 points higher than students who began in the fall. The results from the final model suggested that scores on the math benchmark significantly and positively related to students' intercept. Every one point that students scored above the grand mean on the fall benchmark corresponded with, on average, a .20 point increase on the progress monitoring intercept. This value varied significantly at the school level,  $\chi^2(93) = 194.88, SD = .05, p < .001$ . Math benchmark scores were not a significant predictor of students' linear slope, though this random effect did vary significantly between schools,  $\chi^2(93) = 118.04, SD = .01, p = .04$ . Overall, the model including students' fall math benchmark scores accounted for 75.63% of the student-level variance beyond the unconditional growth model.



## Analysis 2

The model that included students' fall passage reading fluency, reading comprehension, and vocabulary benchmark scores as predictor variables was fit for a substantially reduced sample: Students who had valid data for all three reading measures and were progress monitored in the area of *Number and Operations*.

The results of the unconditional curvilinear model are displayed as *Model 1* in Table 2. The chi-square deviance test suggested the quadratic model fit the data significantly better than the linear growth model,  $\chi^2(7) = 65.07, p < .001$ . Overall, the model suggested that students began, on average, responding correctly to 10.55 out of the 16 possible questions. This value varied significantly between students,  $\chi^2(393) = 996.85, SD = 1.93, p < .001$ , and between schools,  $\chi^2(39) = 135.58, SD = 1.37, p < .001$ . Students progressed, on average, .54 points correct per month, with a deceleration of .05 points squared per month. The linear slope did not vary significantly between students,  $\chi^2(393) = 432.72, p = .08$ , but was retained for further model testing given its theoretical importance. That is, we were interested in the variance of the parameter regardless of its statistical significance. The rate of deceleration did vary between students,  $\chi^2(39) = 471.14, SD = .09, p = .004$ . Both the linear,  $\chi^2(39) = 71.04, SD = .28, p = .002$ , and the quadratic,  $\chi^2(39) = 61.97, SD = .03, p = .011$ , parameters varied significantly between schools. The ICC for the unconditional growth model suggested that 54.30% of students' intercept variance depended on students, while less than 21.55% depended on schools. Additionally, 4.62% of students' linear slope variance depended on students, while 2.29% depended on schools.

Similar to Analysis 1, seasonal control variables were then entered into the model as student-level predictors of their intercept and both slopes. Allowing the dummy variables to vary

randomly between schools resulted in insufficient data to allow for statistical testing of the random effects. Further, the model including the random effects did not fit significantly better than the model without the random effects,  $\chi^2(39) = 18.12, p > .50$ . For parsimony, we thus fixed the random effects for each dummy-vector. The results of this model are displayed in Table 2 as *Model 2*. Both the winter and spring seasonal control start variables were significant predictors of students' intercept. Students who began progress monitoring in the winter scored, on average, 0.94 points higher than students who began in the fall. Similarly, students who began progress monitoring in the spring scored, on average, 1.77 points higher than students who began in the fall. Neither seasonal control variable was a significant predictor of students' linear or quadratic slopes. These variables were thus removed from the model as predictors of either slope prior to further model testing. Overall, the seasonal control model accounted for 15.87% of the student-level variance beyond the unconditional growth model.

Students' fall passage reading fluency, reading comprehension, and vocabulary benchmark scores were next entered into the model grand-mean centered, as predictors of students' intercept and both slopes. All parameters were initially allowed to vary randomly between schools; however, only the effect of vocabulary on students' intercept and quadratic slope varied significantly between schools. We fixed all other random effects, which resulted in the vocabulary random effects no longer varying significantly between schools. All reading random effects were thus fixed. The results of this model are presented as *Model 3* in Table 2.

Overall, the final model fit significantly better than the seasonal control model,  $\chi^2(9) = 194.80, p < .001$ . The final model suggested that students began, on average, responding correctly to 10.64 math questions. This value varied significantly between students,  $\chi^2(387) = 817.10, SD = 1.63, p < .001$ , and between schools,  $\chi^2(40) = 73.28, SD = .52, p < .001$ . Similar

to the unconditional growth model, students progressed, on average, .55 points correct per month, with a deceleration of .05 points squared per month. The linear slope, again, did not vary significantly between students,  $\chi^2(389) = 430.62$ ,  $SD = .39$ ,  $p = .07$ , though the rate of deceleration did,  $\chi^2(389) = 465.42$ ,  $SD = .09$ ,  $p = .005$ . The linear parameter varied significantly between schools  $\chi^2(40) = 74.58$ ,  $SD = .30$ ,  $p = .002$ , as did the quadratic,  $\chi^2(40) = 67.22$ ,  $SD = .04$ ,  $p = .005$ . Both the winter and spring seasonal control start variables remained significant predictors of students' intercept. On average, students who began progress monitoring in the winter scored 1.04 points higher compared to students who began in the fall, while students who began progress monitoring in the spring scored, on average, 2.38 points higher than students who began in the fall.

The results suggested that all three reading predictor variables significantly and positively related to students' mathematics intercept; however, none of the reading variables significantly related to students' linear or quadratic slope, for their math growth. On average, for every point increase in the students' reading comprehension score on the fall benchmark, students' initial mathematics progress monitoring score increased .14 points. For each additional word correctly read on the passage reading fluency measure, students' initial progress monitoring score increased .01 points, and for every point increase on the vocabulary measure, students' initial math score increased by .15 points. Overall, the model including students' performance on the three fall reading benchmark assessments accounted for 29.08% of the student-level variance beyond the unconditional growth model.

### **Discussion**

The results of this study suggest that students who were progress-monitored demonstrated within-year math growth that was statistically observable through a brief 16-item

assessment addressing late elementary skill sets. Though growth in math is likely expected within the school year, when placed within the context of the focus of the current research base, this finding appears to not be trivial. Foegen et al. (2007) found that research has largely focused on growth in reading, and, to a lesser degree, early fluency-based math skills. Specifically, of the 578 reports, dissertations, chapters, and journal articles Foegen and colleagues analyzed, only 32 studies focused on mathematics. Additionally, the statistically significant growth in math we observed was found on measures targeting skills later in the developmental process for elementary students. To date, most research on math progress-monitoring measures has largely addressed early fluency-based skill areas such as addition, subtraction, multiplication, and division, typically monitoring student growth by assessing the change in the number of correct digits over time (L. Fuchs, et al., 2005). Lembke et al. (2012) reported that only three studies (approximately 11% of the total number of studies they analyzed) investigated growth in mathematics beyond the Kindergarten or first grade level. The sparseness of findings focused on growth in mathematics thus suggests a need for greater attention to this topic moving forward. Our study attempts to target this need by investigating the growth in a later elementary grade.

For this reason, we intentionally targeted growth in higher-order (relative to early elementary) mathematics skills by selecting progress monitoring measures were designed to evaluate more complicated and nuanced problems, such as order of operations and fractions, often couched within real-world problem-solving scenarios. Items from the easyCBM *Number and Operations* measure go beyond assessing students' proficiency with fluency-based algorithms, and instead assess students' application of more complex mathematical concepts. And, though assessing more complex skills sets in Grade 4, the measures were quite short, taking on average, less than 12 minutes to administer

([http://www.rti4success.org/tools\\_charts/popups\\_progress/easyCBMMath\\_area.php](http://www.rti4success.org/tools_charts/popups_progress/easyCBMMath_area.php)). Efficient administration of probes of these types is a key feature of progress monitoring within an RTI context, where measures should be short, reliable, and used to make instructional decisions tied to individual student needs (D. Fuchs, et al., 2003). Further, Anderson, Lai, Alonzo and Tindal (2011) showed that the bulk of the math items were targeted at students performing below expectations. These findings are in line with the RTI-implementation guidelines provided by Fuchs and Fuchs (2001), who suggest that students be selected for progress monitoring because they are at risk of not meeting grade-level expectations. Thus, the results of our study further the findings by Anderson and colleagues, suggesting that the progress-monitoring outcome measures used in our study are not only statistically sensitive to detecting students performing below expectations at a single point in time, but also to documenting their improvement over time.

However, it is important to note that the growth we observed was roughly one-fifth of a correct math problem per month (Analysis 1) to half of a problem per month (Analysis 2). In a practitioner/school context, this means that two to five months of monitoring would be required before a teacher would “see” any raw score changes, making any decisions made from “growth” difficult, if not impossible given the 16-point scale in the outcome measure used in this study. Thus, future research should perhaps consider moving from raw score reporting to a scale score so that the statistical growth found on short progress monitoring probes is more readily observable to practitioners as a means to guide instructional decision-making.

Although our results around growth in later elementary developmental skills contributes to the math growth literature beyond the early elementary grades, our findings were not definitive in terms of whether the observed math growth is linear or curvilinear (decelerating rate) within the school year. Whether growth is linear or curvilinear may have important

implications for helping teachers and school leaders to establish short- and long-term goals for students and/or for interpreting interim-formative assessment results in the context of instructional decision-making within RTI. For instance, if students grow at a linear rate from fall to spring, then it is likely sensible to use this information as a means to check the consequential validity of instructional decisions over the course of the year, as students should be progressing at a relatively stable rate over time. On the other hand, if growth decelerates over time, greater growth would be expected early in the school year, a finding that teachers could use to adjust and increase the intensity of instruction (and intervention) earlier or later in the year, while assuming decelerating growth trends over time. Though certainly not the only piece of data used to help establish student academic goals in math, better understanding of the *trajectory* of growth that should be expected is an important piece of information within the constellation of student-based information used for instructional decision-making. Thus, future research on within-year growth in math should seek to more clearly establish whether such growth is linear or not.

Similarly, our finding that students' fall math benchmark scores accounted for 75.63% of the student level variance beyond the unconditional growth model, may suggest the importance of assessing students early in the year as a means of identifying those who may require targeted instruction and intensive intervention. Where students started, in terms of their math benchmark performance was the most significant predictor of how they continued to perform on math measures throughout the year. This finding suggests the need for additional research, not only in terms of perhaps needing more sensitive measures of monitoring progress within year, but also, perhaps, on identifying more effective instructional approaches. With progress measures limited to raw scores on a 16-item scale, the growth we detected would not be interpretable by teachers using the easyCBM measures.

Our findings that students' initial performance (intercept) on the mathematics assessments was only moderately positively related to their performance on the reading benchmark assessments and that reading benchmark score did not add to the variance explained by the growth models provide some evidence to suggest that the attention paid to the principles of UDA during development of the easyCBM mathematics measures was successful in decreasing the construct-irrelevant variance related to reading access skills. This finding has implications for others developing mathematics measures. Some relation between performance on the reading and mathematics benchmark assessments is to be expected, as both assessments likely measure, to some extent, general cognitive processing as well as the specific content areas (math and reading) that are their primary focus.

### **Limitations and Future Research**

Although our study adds to the research base on mathematics performance over time, specifically addressing both a grade level (4<sup>th</sup>) and content area (numbers and operations) that few other studies have addressed, we are left with more questions than answers. Our study does not provide definitive evidence of whether 4<sup>th</sup>-grade student growth in *Numbers and Operations* is best described as linear or curvilinear. In addition, we did not include any variables related to instruction in our models. An important extension to this study would be to include information about the intensity, type, and content of instruction being provided to students. To determine the sensitivity of the easyCBM measures to detect growth in student mathematics skills, it is important to understand whether students have *actually made any progress* over the course of the year. To the extent that more intense, effective, and targeted instruction results in higher performance, future models could take into account instructional variables and their impact on within and across-year growth. Without the inclusion of such instructional variables, however, it

is impossible to decipher our results fully. Perhaps the variation in within-year growth that we found in this study is an indication of the variation in effectiveness of educational practices. This is an area ripe for additional research.



## References

- Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012). The development of CBM vocabulary measures: Grade 4 (Technical Report No. 1211). Eugene, OR: Behavioral Research and Teaching, University of Oregon.
- Alonzo, J., Anderson, D., & Tindal, G. (2009). *IRT analysis of general outcome measures in grades 1-8* (Technical Report No. 0916). Eugene, OR: Behavioral Research and Teaching, University of Oregon.
- Alonzo, J., Lai, C. F., & Tindal, G. (2009). The development of K-8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 4 (Technical Report No. 0903). Eugene, OR: Behavioral Research and Teaching, University of Oregon.
- Alonzo, J., Liu, K., & Tindal, G. (2007). Examining the technical adequacy of reading comprehension measures in a progress monitoring assessment system (Technical Report No. 41). Eugene, OR: Behavioral Research and Teaching, University of Oregon.
- Alonzo, J., & Tindal, G. (2007). The development of word and passage reading fluency measures in a progress monitoring assessment system (Technical Report No. 40). Eugene, OR: Behavioral Research and Teaching, University of Oregon.
- Anderson, D., Alonzo, J., & Tindal, G. (2010a). easyCBM mathematics criterion related validity evidence: Oregon state test (Technical Report No. 1011). Eugene, OR: Behavioral Research and Teaching, University of Oregon.
- Anderson, D., Alonzo, J., & Tindal, G. (2010b). easyCBM mathematics criterion related validity evidence: Washington state test (Technical Report No. 1010). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

- Anderson, D., Lai, C., Alonzo, J., & Tindal, G. (2011). Examining a grade-level math CBM designed for persistently low-performing students. *Educational Assessment, 16*, 15-34. doi: 10.1080/10627197.2011.551084
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice, 5*(1), 7-74.
- Calhoon, M. (2008). Curriculum-based measurement for mathematics at the high school level: what we do not know and what we need to know. *Assessment for Effective Intervention, 33*(234-239).
- Foegen, A., & Deno, S. L. (2001). Identifying growth indicators for low-achieving students in middle school mathematics. *Journal of Special Education, 35*(1), 4-16.
- Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *The Journal of Special Educatio, 41*(2), 121-139. doi: 10.1177/00224669070410020101
- Fuchs, D., & Fuchs, L. (2001). Responsiveness-to-intervention: A blueprint for practitioners, policymakers, and parents. *Teaching Exceptional Children, 38*(1), 57-61.
- Fuchs, D., & Fuchs, L. (2006). Introduction to response to intervention: What, why, and how valid is it? *Reading Research Quarterly, 41*, 93-99. doi: 10.1598/RRQ.41.1.4
- Fuchs, D., Mock, D., Morgan, P., & Young, C. (2003). Responsiveness-to-intervention: Definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research & Practice, 18*(157-171). doi: 10.1111/1540-5826.00072
- Fuchs, L., Fuchs, D., & Courey, S. (2005). Curriculum-based measurement of mathematics competence: From competence to concepts and applications to real life problem-solving. *Assessment for Effective Intervention, 30*(2), 33-46. doi: 10.1177/073724770503000204

- Lembke, E., Hampton, D., & Beyers, S. J. (2012). Response to intervention in mathematics: Critical elements. *Psychology in the Schools, 49*, 257-272. doi: 10.1002/pits.21596
- Nese, J. F. T., Lai, C. F., Anderson, D., Jamgochian, E. M., Kamata, A., Saez, L., Park, B. J., Alonzo, J., Tindal, G. (2010). Technical adequacy of the easyCBM mathematics measures: Grades 3-8, 2009-2010 Version (Technical Report No. 1007). Eugene, OR: Behavioral Research and Teaching, University of Oregon.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Second ed.). Thousand Oaks, CA: Sage.
- Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., & du Toit, M. (2011). HLM Version 7: Scientific Software International, Inc.
- Saez, L., Park, B. J., Nese, J. F. T., Jamgochian, E. M., Lai, C. F., Anderson, D., . . . Tindal, G. (2010). Technical adequacy of the easyCBM reading measures (Grades 3-7), 2009-2010 Version (Technical Report No. 1005). Eugene, OR: Behavioral Research and Teaching, University of Oregon.
- Stevens, J. J. (1995). Confirmatory factor analysis of the Iowa tests of basic skills. *Structural Equation Modeling, 2*, 214-231. doi: 10.1080/10705519509540010
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved November 2, 2011, from the World Wide Web: <http://www.cehd.umn.edu/nceo/OnlinePubs/Synthesis44.html>.

Table 1

*Fixed and random effects for math progress monitoring growth (Analysis 1)*

Fixed Effects	Model 1		Model 2		Model 3	
	Coefficient	SE	Coefficient	SE	Coefficient	SE
Intercept, $\beta_{00}$	10.68	0.14	10.38	0.16	10.69	0.09
Winter start, $\beta_{01}$			0.66	0.26	0.56	0.12
Spring start, $\beta_{02}$			2.36	0.60	1.79	0.25
MATH, $\beta_{03}$					0.25	0.01
Linear Slope, $\beta_{10}$	0.27	0.05	0.20	0.02	0.20	0.02
Winter start, $\beta_{11}$			- 0.03 <sup>†</sup>	0.06		
Spring start, $\beta_{12}$			0.17 <sup>†</sup>	0.09		
MATH, $\beta_{13}$					< 0.01 <sup>†</sup>	< 0.01
Quadratic Slope, $\beta_{20}$	-0.01 <sup>†</sup>	0.01				
Random effects	Estimate		Estimate		Estimate	
Within student, $e_{i(jk)}$	3.10		3.18		3.19	
Btwn student intercept, $r_{0j}$	3.99		3.67		0.97	
Btwn student linear, $r_{1j}$	0.15		0.03		0.02	
Btwn student quadratic, $r_{2j}$	< 0.01 <sup>†</sup>					
Btwn school intercept, $u_{00}$	1.29		1.19		0.36	
Btwn school benchmark on intercept, $u_{03}$					< 0.01	
Btwn school linear, $u_{10}$	0.09					
Btwn school benchmark on linear, $u_{13}$					< 0.01	
Btwn school quadratic, $u_{20}$	< 0.01					
Model deviance ( $df$ )	42335.11 (7)		42329.77 (20)		40450.49 (16)	

<sup>†</sup> Coefficient is *not* significant,  $p > .05$ ; All other values significant,  $p < .05$ .

Table 2

*Fixed and random effects for math progress monitoring growth (Analysis 2)*

Fixed Effects	Model 1		Model 2		Model 3	
	Coefficient	SE	Coefficient	SE	Coefficient	SE
Intercept, $\beta_{00}$	10.55	0.27	10.35	0.25	10.64	0.17
Winter start, $\beta_{01}$			0.94	0.49	1.04	0.38
Spring start, $\beta_{02}$			1.77	0.57	2.38	0.26
MCRC, $\beta_{03}$					0.14	0.03
PRF, $\beta_{04}$					0.01	< 0.01
VOC, $\beta_{05}$					0.15	0.05
Linear Slope, $\beta_{10}$	0.53	0.08	0.54	0.09	0.55	0.08
Winter start, $\beta_{11}$			0.16 <sup>†</sup>	1.89 <sup>†</sup>		
Spring start, $\beta_{12}$			1.17 <sup>†</sup>	1.13		
MCRC, $\beta_{13}$					-0.02 <sup>†</sup>	0.02
PRF, $\beta_{14}$					< 0.01 <sup>†</sup>	< 0.01
VOC, $\beta_{15}$					0.01 <sup>†</sup>	0.02
Quadratic Slope, $\beta_{20}$	-0.05	0.01	-0.05	0.01	-0.05	0.01
Winter start, $\beta_{21}$			-0.04 <sup>†</sup>	0.05 <sup>†</sup>		
Spring start, $\beta_{22}$			< 0.01 <sup>†</sup>	0.62 <sup>†</sup>		
MCRC, $\beta_{23}$					0.01 <sup>†</sup>	< 0.01
PRF, $\beta_{24}$					< 0.01 <sup>†</sup>	< 0.01
VOC, $\beta_{25}$					< 0.01 <sup>†</sup>	< 0.01
Random effects	Estimate		Estimate		Estimate	
Within student, $e_{i(jk)}$	3.14		3.14		3.14	
Btwn student intercept, $r_{0j}$	3.73		3.78		2.65	
Btwn student linear, $r_{1j}$	0.16 <sup>†</sup>		0.15 <sup>†</sup>		0.15 <sup>†</sup>	
Btwn student quadratic, $r_{2j}$	0.01		0.01		0.01	
Btwn school intercept, $u_{00}$	1.88		1.12		0.27	
Btwn school linear, $u_{10}$	0.08		0.08		0.09	
Btwn school quadratic, $u_{20}$	< 0.01		< 0.01		< 0.01	
Model deviance ( $df$ )	12123.14 (16)		12113.38 (18)		11918.58 (27)	

<sup>†</sup> Coefficient is *not* significant,  $p > .05$ ; All other values significant,  $p < .05$ .

The interface displays a math problem on the left and three answer choices on the right. The number line has three points: 1.25, a point indicated by a downward arrow, and 1.75. The question asks "Which is here?". The answer choices are 1.60, 1.70, and 1.50. The choice 1.50 is selected and highlighted in blue. Navigation buttons "Back" and "Next" with a right arrow are at the bottom right, along with a speaker icon for audio.

1.25      1.75

Which is here?

☐ 1.60

☐ 1.70

☒ 1.50

Back      Next →

Speaker icon

Figure 1. Sample math item from the Grade 4 *Numbers and Operations* progress monitoring measure with the correct answer highlighted.

## Appendix A

Full *A Priori* Conditional Math BM Predictor Model (Analysis 1)

$$PMScore_{tij} = \pi_{0ij} + \pi_{1ij}(time) + \pi_{2ij}(time^2) + e_{tij}$$

$$\pi_{0ij} = \beta_{00j} + \beta_{01j}(WintStart) + \beta_{02j}(SprStart) + \beta_{03j}(FallMathBM) + r_{0ij}$$

$$\pi_{1ij} = \beta_{10j} + \beta_{11j}(WintStart) + \beta_{12j}(SprStart) + \beta_{13j}(FallMathBM) + r_{1ij}$$

$$\pi_{2ij} = \beta_{20j} + \beta_{21j}(WintStart) + \beta_{22j}(SprStart) + \beta_{23j}(FallMathBM) + r_{2ij}$$

$$\beta_{00j} = \gamma_{000} + u_{00j}$$

$$\beta_{01j} = \gamma_{010} + u_{01j}$$

$$\beta_{02j} = \gamma_{020} + u_{02j}$$

$$\beta_{03j} = \gamma_{030} + u_{03j}$$

$$\beta_{10j} = \gamma_{100} + u_{10j}$$

$$\beta_{11j} = \gamma_{110} + u_{11j}$$

$$\beta_{12j} = \gamma_{120} + u_{12j}$$

$$\beta_{13j} = \gamma_{130} + u_{13j}$$

$$\beta_{20j} = \gamma_{200} + u_{20j}$$

$$\beta_{21j} = \gamma_{210} + u_{21j}$$

$$\beta_{22j} = \gamma_{220} + u_{22j}$$

$$\beta_{23j} = \gamma_{230} + u_{23j}$$

Full *A Priori* Conditional Reading BM Predictor Model (Analysis 2)

$$PMScore_{tij} = \pi_{0ij} + \pi_{1ij}(time) + \pi_{2ij}(time^2) + e_{tij}$$

$$\pi_{0ij} = \beta_{00j} + \beta_{01j}(WintStart) + \beta_{02j}(SprStart) + \beta_{03j}(FallBM\_PRF) \\ + \beta_{04j}(FallBM\_MCRC) + \beta_{05j}(FallBM\_VOC) + r_{0ij}$$

$$\pi_{1ij} = \beta_{10j} + \beta_{11j}(WintStart) + \beta_{12j}(SprStart) + \beta_{13j}(FallBM\_PRF) \\ + \beta_{14j}(FallBM\_MCRC) + \beta_{15j}(FallBM\_VOC) + r_{1ij}$$

$$\pi_{2ij} = \beta_{20j} + \beta_{21j}(WintStart) + \beta_{22j}(SprStart) + \beta_{23j}(FallBM\_PRF) \\ + \beta_{24j}(FallBM\_MCRC) + \beta_{25j}(FallBM\_VOC) + r_{2ij}$$

$$\beta_{00j} = \gamma_{000} + u_{00j}$$

$$\beta_{01j} = \gamma_{010} + u_{01j}$$

$$\beta_{02j} = \gamma_{020} + u_{02j}$$

$$\beta_{03j} = \gamma_{030} + u_{03j}$$

$$\beta_{04j} = \gamma_{040} + u_{04j}$$

$$\beta_{05j} = \gamma_{050} + u_{05j}$$

$$\beta_{10j} = \gamma_{100} + u_{10j}$$

$$\beta_{11j} = \gamma_{110} + u_{11j}$$

$$\beta_{12j} = \gamma_{120} + u_{12j}$$

$$\beta_{13j} = \gamma_{130} + u_{13j}$$

$$\beta_{14j} = \gamma_{140} + u_{14j}$$

$$\beta_{15j} = \gamma_{150} + u_{15j}$$

$$\beta_{20j} = \gamma_{200} + u_{20j}$$

$$\beta_{21j} = \gamma_{210} + u_{21j}$$

$$\beta_{22j} = \gamma_{220} + u_{22j}$$

$$\beta_{23j} = \gamma_{230} + u_{23j}$$

$$\beta_{24j} = \gamma_{240} + u_{24j}$$

$$\beta_{25j} = \gamma_{250} + u_{25j}$$