Journal of Educational Administration and Policy, Vol. 2 (December 2017), 33-46
ISSN 2508-5573 | http://doi.org/10.22553/keas/2017.2.1.33

Journal of Educational Administration and Policy

http://www.keas.or.kr/en/

A Validity Argument for a Mathematics Curriculum-Based Measure: Implications for Response to Intervention Decision-Making

B. Jasmine Park[a] · Daniel Anderson[b] · Gerald Tindal[b] · Julie Alonzo[b]

[a]*American Institutes for Research*
[b]*University of Oregon*

A B S T R A C T

Within a response to intervention (RTI) framework, many schools initially group students into tiers on the basis of normative achievement. Using a screener and benchmarks with curriculum-based measurement (CBM), students are classified as being academically "at-risk" or not. In the current study, we present a validity argument for the use of a mathematics CBM as a classification tool within RTI and explore the relation between a fall CBM administration and state accountability test results using a large sample in Oregon (located in the Pacific Northwest in the United States) through regression analyses. Our analyses indicate a strong relation between easyCBM® and the state achievement tests, which suggests that accurate screening of students using CBM can help teachers and administrators make more informed decisions for both instruction and school resource allocation. For more successful RTI, educators should implement effective and efficient decision-making processes based on assessment data, and researchers should continue exploring ways to improve the classification accuracy of CBM and incorporate instructional effects.

Keywords: Curriculum-based measurement, response to intervention, classification accuracy, test validity, decision making

Introduction

Response to intervention (RTI) is a decision-making framework used to identify and monitor the progress of students at-risk for low achievement using curriculum-based measurement (CBM) probes. The effectiveness of the instructional program can then be evaluated and modified based on the students' progress on the CBM probes. The purpose of this paper is to present evidence to add to the validity argument for the use of a mathematics CBM as a classification tool within an RTI framework. The validity argument follows the logic outlined by Kane (1992, 2006). To evaluate validity, we examine the underlying assumptions of the inferences made from the assessment results and present lines of evidence to confirm that the assumptions are met.

We argue that when a CBM is used to classify students, it essentially becomes a placement test. Although in practice CBM is only one indicator of potentially many used to inform placement decisions, it is often the primary indicator and thus warrants particular scrutiny. Kane (2006) states that when building an interpretive argument for a placement test, there are four basic inferences: scoring, generalization, extrapolation, and decision. We investigate each of these inferences and their underlying assumptions, but begin by first clarifying which students are intended

* Corresponding Author: Bitnara Jasmine Park
American Institutes for Research: 1000 Thomas Jefferson Street NW. Washington, D.C. 20007 U.S.A.
Tel: +1-202-403-5704
E-mail: bpark@air.org

to be identified with the CBM benchmark screener.

This study adds to the published validity evidence for the extrapolation and decision inferences for the use of the fall easyCBM® mathematics measures in grades 6-8 within an RTI framework using a large samples from three districts in Oregon, located in the Pacific Northwest of the United States. Robust validity evidence of CBM is an essential factor for successful RTI implementation because student placement for tiered instruction based on the CBM assessment data has implications not only for instructional planning and practice, but also for school resource allocation. Valid use of CBM can help educators make informed decisions to maximize efficiency of limited school resources by providing the right kind of support to target students' unique needs.

The Target Student Group

The purpose of RTI is to provide support for students at-risk for low achievement by grouping them into a "tier", often called Tier 2, and providing an instructional intervention (D. Fuchs & Fuchs, 2001). In this study, we examine easyCBM®, an assessment tool designed for use within RTI for two different but complementary purposes: (a) identification of at-risk students with a benchmark screening assessment, and (b) monitoring their progress after Tier 2 placement with multiple forms of equivalent difficulty. We focus here only on the validity of using the fall easyCBM® mathematics measure to classify students as at-risk in grades 6-8, not on the progress monitoring measures.

Within RTI, some portion of students will always be identified as at-risk. The size of this group depends on the school's or district's RTI model, but students are typically identified using a normative achievement level on a benchmark screener (L. Fuchs & Fuchs, 2007). Students scoring below this normative cut point are placed in Tier 2, while those scoring above the cut point receive instruction from the general education curriculum (Tier 1). Although educators set the normative cut point for Tier 2 placement, it is not the cut point itself that defines whether a student is at-risk or not. For instance, if a student performs at the 18th percentile in a school where the cutoff for Tier 2 placement is the 20th percentile, he or she will be placed in Tier 2. If this same student moves to a school where the cutoff is the 15th percentile, he or she would no longer be identified as at-risk. The student has not changed, yet the identification of need has changed (and the subsequent resources provided to that student). A similar scenario is possible in the other direction, where a student moves from being identified as "on-track" to at-risk. When evaluating whether a measure accurately identifies students as at-risk, researchers must determine that it does so relative to students' characteristics rather than the RTI protocols surrounding a school.

Although imperfect, perhaps the best criterion for establishing if students qualify for Tier 2 placement in the beginning of the year is their performance on the state test at the end of the year in grade levels where a state assessment is given. Using state test performance as the criterion, it is possible to conduct a retrospective analysis to examine if the screening measure administered in the fall accurately predicted students' performance on the state test in the spring, regardless of whether the student was placed in Tier 1 or Tier 2 for instruction. This approach targets the characteristics of the student (knowledge and skills displayed on the state test) and not the cut points used to access Tier 2.

Using state test performance as the criterion, however, does not account for the instructional practices occurring in the classroom or school between the interval of the screener and the state test administrations. It is therefore possible that the overall predictive accuracy of the screener is weakened, as it is unknown whether students who performed low on the fall screener received an intervention and Tier 2 placement, or if they remained in Tier 1. If the intervention and Tier 2 placement were highly effective, students who performed low on the screener in the fall but received Tier 2 interventions might perform quite differently from those who performed poorly on the screener in the fall but did not receive Tier 2 interventions. If tier placement was not effective, however, the overall predictive accuracy of the screener would be similar for both groups. Not accounting for tier placement likely either weakens the predictive accuracy or results in the same predictive accuracy as if tier placement were accounted for. Thus, if the screener remains an adequate predictor of state test performance without accounting for instructional effects, it would increase the validity evidence for use as a classification tool.

Knowing which students are at-risk for failing the state test early in the year could be of substantial benefit.

The educators could use an "at-risk for failing the state test" index as part of their decision making for initial tier placement. Of course, CBMs should never be the only indicator used for identification of risk, and state test performance should not be the only consideration of whether the student is at-risk or not. But because CBMs often serve as the primary source of information within an RTI framework, and many educators are concerned with state test performance, high predictive validity of state test performance is an important characteristic of a classification tool.

Inferences from CBM Benchmark Screeners

If we accept that students' state test performance at the end of the year is a reasonable criterion for establishing whether students needed Tier 2 placement in the beginning of the year, we can then begin to examine other inferences drawn from the test results, and the underlying assumptions of those inferences. In what follows, we discuss the four inferences of placement tests outlined by Kane (2006), and the assumptions involved with each of them. We then discuss the evidence needed to confirm that these assumptions are met, and the research that has been conducted with mathematics CBMs in each area. We conclude each section with a discussion of the validity evidence collected to date for the current CBM under investigation, easyCBM®. We are primarily concerned with the extrapolation and decision inferences in this study, so we give preference to these areas over the others.

Scoring. When a CBM is administered, the student receives a score. Two primary assumptions associated are with score inferences: (a) the score is appropriate, and (b) the score is consistent (Kane, 2006). Although CBM is designed for classroom use, all procedures and scoring protocols are standardized. According to Deno (2003), this standardization increases the reliability of the measures—a requisite condition of validity (Kane, 1992)—while also increasing scoring consistency. In mathematics CBM, some researchers have assessed computational fluency by scoring the number of correct digits (i.e., L. Fuchs et al., 2007). In the upper grades, as the curriculum becomes more demanding, simple raw scores such as the number of correct digits are likely to become less useful. Items assessing students' knowledge of the communicative property, for instance, or of geometric relations, would require different scoring algorithms. Typically, these items are presented in a multiple choice format and scored dichotomously correct/incorrect (e.g., Espin et al., 2009).

To enhance the standardization of administration and scoring protocols, easyCBM® was developed to be administered by a computer. Thus, at each testing occasion the student is presented with the items in the same way with all items scored dichotomously (correct/incorrect). Computer administration and scoring helps to ensure that all students are presented with the test in the same way, regardless of school setting, which increases the scoring consistency.

Generalization. The generalizability of a test score is a function of observation and domain sampling (Kane, 2006). When a teacher administers a CBM, the student is being assessed on a limited number of items with two assumptions: (a) the items in the test are representative of the entire universe of items, and (b) a sufficient number of items are present to control for sampling error. Reliability or generalizability studies provide empirical evidence that a test meets, or does not meet, the generalization inference assumptions.

Foegen, Jiban, and Deno (2007) conducted an extensive review of CBM mathematics research literature and reported the reliability under various conditions. Because CBMs are designed to measure students' growth using multiple forms (Deno, 2003), a substantial amount of reliability research has focused on alternate form comparability. When CBM is used as a classification tool, however, alternate form reliability may be less important. Rather, the internal consistency of the measure should become the focus, along with the consistency with which students receive a particular score (test-retest). Foegen and colleagues (2007) reported internal consistency statistics from five studies, ranging from .94 to .98, with two other studies providing more modest coefficients (.60 and .83). Of the seven studies that addressed test-retest reliabilities, coefficients were generally in the .70 to .80 range, but included coefficients ranging from a low of .48 to a high of .97. Three other studies (Christ, Johnson-Gros, & Hintze, 2005; Christ & Vining, 2006; Hintze, Christ, & Keller, 2002) used generalizability theory to examine sources of error in a mathematics CBM under a variety of conditions (domain breadth, skill, and form). The results of these studies were generally mixed, but indicated that the most variance was associated with the student when

Table 1. National Council of Teachers of Mathematics (NCTM) Focal Point – Grades 6-8

Grade	Focal point 1	Focal point 2	Focal point 3
6	Number and operations	Algebra	Number and Operations Rate/Ratio
7	Number and Operations and Algebra and Geometry	Measurement and Geometry and Algebra	Number and Operations and Algebra
8	Algebra	Geometry, and Measurement	Data Analysis and Number and Operations and Algebra

a stratified item sampling design was used (Christ & Vining, 2006), and forms targeted a single skill (Hintze et al., 2002).

In the current study, we document the internal consistency of the test to focus on the sample of items most highly related to an eventual criterion of a state test and downplay the dependence of the placement on the cut scores (which would involve decision consistency, not item consistency). Items on the easyCBM® mathematics measures were developed using principles of Universal Design for Assessment to reduce irrelevant cognitive and language complexity. All easyCBM® forms were constructed to address the Mathematics Focal Point Standards as outlined by the National Council of Mathematics Teachers (See Table 1). We developed items in these focal points, calculated item difficulty using a Rasch model, and created alternate forms by sampling items equally across the different focal points and difficulties. Nese and colleagues (2010) report internal consistency with Cronbach's alpha for the fall benchmark in grades 6 to 8 as .87, .89, and .89 respectively, with split-half reliability estimates ranging from .73 to .82.

Extrapolation. If the assumptions underlying the score and generalization inferences have been met, an extrapolation can be made to infer a level of skill (Kane, 2006). An extrapolation inference assumes that the tasks within the test match the curriculum used in the classroom over a period of time, and that the test does not require skills irrelevant to students' classroom success. Evidence that the extrapolation assumption is met can take the form of documentation of the test development process (i.e., including content experts in the item writing), or empirical investigations of the relation between the measure of interest and existing measures assessing the same domain.

Examining the extrapolation inference with an existing measure requires the additional assumption that the criterion measure used is an adequate measure of students' abilities, and that both measures target the same domains. This assumption is likely met when the state test is used as the criterion, as items within state tests are written to target the instructional standards of the state. When a criterion measure other than the state test is used (e.g., Helwig, Anderson, & Tindal, 2002), the extrapolation validity argument is weakened. Measures other than the state tests may target skills irrelevant to students' classroom success in a particular state, while *not* addressing skills pertinent to students' classroom success. Espin and associates (2009) examined the relation between a mathematics CBM and three criterion measures, including the Minnesota state test, and found that the CBM typically correlated stronger with the other criterion measures than with the state test. Thus, estimates of the relation between the measure and classroom instruction could be inflated when using a criterion measure other than the state test.

A considerable amount of research has documented the relation between CBMs and state test performance in reading (Burke, Hagan-Burke, Kwok, & Parker, 2008; Fewster & Macmillan, 2002; Goffreda, Diperna, & Pedersen, 2009), but relatively little has been published in mathematics (Chard et al., 2005; Shapiro, Keller, Santoro, & Hintze, 2006). Studies by Espin et al. (2009), Shapiro et al. (2006), and Keller-Margulis, Shapiro, and Hintze (2008) are three notable exceptions. Espin and colleagues (2009) examined the relation of only computation-based probes with Minnesota's state test in grades 3 and 5. Using fall and winter benchmark data, the authors found correlations ranging from .44 (grade 3 – winter) to .55 (grade 5 – fall). Shapiroand and others (2006) examined the relation between both computation and concept/ application-based probes with the Pennsylvania state test, then conducted receiver operating characteristics (ROC) curve analyses to establish optimal cut scores. The authors found correlations ranging from .07 (computation, grade 5 – fall) to .64 (concepts/applications, grade 3 – spring). In a subsequent investigation of Pennsylvania elementary students, Keller-Margulis and colleagues (2008) explored the relation between scores on CBM mathematics

computation and concepts/applications probes and scores on the Pennsylvania state assessment. The authors found moderate correlations between CBM mathematics benchmark scores and subsequent state assessment scores ranging from .52 to .66 in grade 2, .14 to .58 in grade 4, .27 to .59 in grade 1, and from .40 to .49 in grade 3. The authors also found strong evidence of diagnostic accuracy of the CBM mathematics assessments at predicting student success on the state assessment one and two years later.

The results of these studies provide some preliminary evidence that mathematics CBMs constructed to assess students' computational fluency or application of concepts may meet the extrapolation inference assumptions, but are limited in their findings. None of these studies explored the relation between the CBM and the state test beyond simple bivariate correlations or optimal cut scores. Additionally, the sample used by Espin et al. (2009) and Shapiro et al. (2006) were relatively small. The Espin et al. (2009) sample was from one urban and one rural Midwestern school district (grade 3, n = 111; grade 5, n = 130), while Shapiro et al. (2006) used a stratified random sample from two districts in Pennsylvania with diverse demographic populations (n = 119 – 206, varying by grade level).

In its development, easyCBM® represents a paradigmatic shift from other mathematics CBMs. Rather than targeting items at the skills of computational fluency or concept application, items were written to a more global outcome, the National Council of Teachers of Mathematics (NCTM) focal point standards, displayed in Table 1. Nese and colleagues (2010) conducted an alignment study to examine the match between easyCBM® items and the NCTM standards in grade K-8. In this study, teacher content experts rated items as *not linked*, *vaguely linked*, *somewhat linked*, or *directly linked* to standards. Items were considered linked if the overall rating of an item was *somewhat* or *directly* linked to the standard objectives. The percent of items rated as *linked* varied from 96% in grade 6, to 83% in grade 7, and 63% in grade 8.

However, to infer that the items are representative of classroom instruction requires the further assumptions that the NCTM standards align with the state standards and that classroom instruction is aligned with these standards. For Oregon, at the time of the study, a simple examination of the state standards relative to the NCTM standards highlights their close association. This association is not surprising given that the NCTM focal point standards were used as a guide during standard writing in the state (Oregon Department of Education, 2008).

Given that a primary purpose of this study was to investigate the relation between easyCBM® and the state tests in Oregon, this linkage of items to standards and standards to instruction is critical. The strong relation between the CBM and state tests and the content standards in each state increases the credibility of evidence for the extrapolation inference – that easyCBM® measures content relevant to classroom instruction.

Decision. If all the assumptions for scoring, generalization, and finally, the extrapolation inference have been met, we can begin to draw conclusions about a student's level of skill upon which to base decisions. When classifying students, three assumptions must first be met: (a) the students' performance later in the school year is dependent on the skill level early in the school year, (b) students with a low level of skill are not likely to succeed without remedial placement, and (c) students with a high level of skill would not also benefit from the remedial placement (Kane, 2006). Within RTI, however, the validity of the decision should be viewed through the lens of both Kane's framework and the practical resources available.

Part of the reason schools choose different Tier 2 cut points is to control access and need for services as the amount of available resources and the level of support needs vary across schools. For example, a lower cut point (in percentile rank metrics) would lower the number of students considered at risk while a higher cut point (again, in percentile rank metrics) would increase the number of students identified as at risk. From the benchmarking data, increasing the normative cut point (i.e., from the 20th to the 25th percentile) can reduce false negative classifications. However, a higher normative cut point, in turn, requires greater resource allocation, as more students become classified as at-risk. Such a decision may be a valid option for schools with ample resources (i.e., even if some students are misclassified, it assures additional assistance for the majority who are in need). In schools with limited resources, however, educators must be equally concerned with false positives, as they must discriminate between students who might benefit from additional attention and students who *need* additional attention. Thus, the third assumption from Kane's (2006) framework may be less important for RTI classification decisions than for typical placement tests.

Little research has been conducted on RTI classification decisions or on the tendency of screening measures to over- or under-identify students as at-risk. In the current study, we examine the relation between the easyCBM® mathematics tests in grades 6-8 and the mathematics portion of the state test in Oregon. Assumptions underlying the extrapolation inference are investigated by examining the relation between the raw scores on easyCBM® and the state test. Assumptions underlying the decision inference are investigated by using easyCBM® scores to predict students into *passing/not passing* categories on the state test. We compare the predicted classification to the observed classification. This investigation does not examine cut scores, but rather uses each student's score to predict classification within a regression framework controlling for demographic characteristics. This study adds to the literature on the validity of using mathematics CBM as a screening tool to classify students into tiers within an RTI framework.

Methods

According to Kane (1992), potential sources of systematic error should be examined empirically. One plausible source of systematic error that is most prominent in the current study is student demographic data (i.e., perhaps students who are English language learners function systematically differently on easyCBM® from those who are not). We controlled for these by including demographic data in all analyses as control variables.

Setting and Participants

Three districts in Oregon participated in this study. Data were collected from the fall administrations of the easyCBM® benchmark assessment during the 2009-2010 school year. The demographics and the size of the Oregon sample are reported by grade level in Table 2. The sample was one of convenience, as each district had adopted easyCBM® as their RTI tool and were willing to share their data with the researchers. All students present on the days of testing in each district were included in the sample. District 1 and District 2 were located in mid-sized towns, while District 3 was part of a large metropolitan area. The three districts ranged in size, with about 800, 1300, and 1,600 students per grade level in District 1, 2, and 3, respectively.

Each district reported student ethnicity in six categories: American Indian/Alaskan Native, Asian, Black, Hispanic, Native Hawaiian/Pacific Islander, or White. Because student ethnicity was included as a control variable, and as not a primary variable of interest, the six ethnic categories were collapsed into two prior to analysis: Non-minority students (White) and Minority students (all remaining ethnic categories). In District 1, 75% of students were identified as White. In Districts 2 and 3, 73% and 59% of students were identified as White, respectively. Approximately 55% of students in the District 1 sample qualified for free or reduced price meals (FRL; a proxy for poverty) while approximately 40% qualified in District 2. These data were not available for District 3. Approximately 16%, 15%, and 13% of students received special education services in Districts 1-3, respectively.

Predictor Variables

The predictor variables used in this study included student demographics and scores from the easyCBM® fall benchmark mathematics assessments. The easyCBM® benchmark mathematics assessments were comprised of 45 multiple-choice mathematics items aligned to the NCTM focal points displayed in Table 1, with approximately 15 aligned to each focal point.

Table 2. Demographics

Grade	*Sample Size*	English language learner (%)	Students with disability (%)	Ethnic minority students (%)	Female (%)	Students met state standards (%)
6	3248	6.59	15.49	34.79	50.28	78.69
7	3055	4.48	12.80	32.96	50.38	83.01
8	3084	4.73	12.78	34.73	47.57	71.92

Criterion Measure

The mathematics portion of the 2009-2010 administration of the Oregon Assessment of Knowledge and Skills (OAKS) was used as the dependent variable. OAKS is a computer-adaptive assessment administered during the spring of 2010 (Oregon Department of Education, 2010). Students were allowed up to three attempts on the OAKS throughout the school year, with their highest score retained for accountability purposes. Students' highest OAKS scores were used in all analyses. The OAKS standard achievement scores are based on Rasch Units, a continuous scale ranging from 0 to infinity (RIT; Oregon Department of Education, 2010). According to the Oregon Department of Education (2010), most scores range from 150-300. Cut scores for the *meets* category during the 2009-2010 school year were 221, 226, and 230 for grades six, seven, and eight, respectively. Means and standard deviations for the OAKS and easyCBM® scores for our sample are reported in Table 3. Student performance on the OAKS is classified into four levels for accountability purposes: *does not meet*, *nearly meets*, *meets*, and *exceeds*. However, because we focus on the accuracy of classification between students passing and not passing the OAKS, the categories were collapsed into dichotomous *does not meet/nearly meets* and *meets/exceeds* categories for analysis.

The assumptions of the generalizability inference for the OAKS were investigated by examining the standard error of measurement at different levels of students' estimated ability, and by specific student subgroups (i.e., ethnic group, special education placement, etc.; Oregon Department of Education, 2006-2007). With the exception of the tails of the distribution, the standard error was reported to be quite consistent regardless of the student subgroup investigated. The state provided evidence to support the assumptions of the extrapolation inference in the technical manual, described the test developmental process, by which items were written to align with the state academic standards. Content experts and multiple review committees then judged the alignment of the items to the standards to be adequate. The decision inference assumptions were examined by investigating the reliability of the achievement classifications. Classification reliability was investigated by examining the standard error around the cut point. According to the OAKS technical manual (Oregon Department of Education, 2006-2007), the reliability coefficients ranged from 84-99%, with most above 90%.

Analyses

We conducted two separate analyses—sequential multiple linear regression (MR) and simultaneous logistic regression (LR)—to examine the relation between the mathematics portion of the fall easyCBM® benchmark in grades 6-8 and the state assessment. MR was conducted to examine the unique relation of each independent variable with state test performance, while LR was conducted to examine the relative importance of each predictor variable in predicting state test classification (passing/not passing). Each grade was run separately.

In both analyses, the independent variables included student demographic data and the fall easyCBM® total score. Four demographic variables were included in each analysis: minority status, special education placement (SPED), sex, and English language learner (ELL) status, with all demographic variables coded dichotomously (0 and 1). In the sequential MR analyses, the student demographic data were entered into the first block, while score on the fall easyCBM® measure was entered into the second block. Sequential MR allowed for an evaluation of the change in total variance accounted for by the model when easyCBM® was included, after accounting for demographic features. We used Tolerance and Variance Inflation Factor (VIF) statistics to address concerns related to multicollinearity. Based on our analyses, the data did not exhibit substantial multicollinearity: We report the range of each statistic for each state in the notes section of Table 4.

Table 3. Descriptive statistics

Variable	Grade 6			Grade 7			Grade 8		
	n	Mean	*SD*	*n*	Mean	*SD*	*n*	Mean	*SD*
State test	3248	227.44	9.73	3055	233.63	9.38	3084	235.51	10.75
easyCBM® Fall total	3248	29.87	7.04	3055	29.41	8.09	3084	28.62	8.15

Table 4. Multiple Regression Coefficients

Grade/variables	*b (SE)*	*β*	*r* (unique variance)
6			
Intercept	197.54 (.54)*		
Minority	-0.94 (.23)*	-.05	-.19 (0.2%)
SPED	-2.41 (.30)*	-.09	-.32 (0.7%)
Female	-0.40 (.21)	-.02	-.05 (0.0%)
ELL	-1.29 (.45)*	-.03	-.21 (0.1%)
Fall total	1.03 (.02)*	.75	.79 (47.9%)
7			
Intercept	207.76 (.47)*		
Minority	-0.54 (.23)*	-.03	-.18 (0.1%)
SPED	-1.55 (.32)*	-.06	-.31 (0.3%)
Female	-0.26 (.20)	-.01	-.02 (0.0%)
ELL	-0.78 (.52)	-.02	-.20 (0.0%)
Fall total	0.90 (.01)*	.78	.80 (49.7%)
8			
Intercept	207.78 (.50)*		
Minority	-1.38 (.25)*	-.06	-.21 (0.3%)
SPED	-2.11 (.36)*	-.07	-.33 (0.4%)
Female	-1.21 (.23)*	-.06	-.05 (0.3%)
ELL	-0.79 (.56)	-.02	-.19 (0.0%)
Fall total	1.02 (.02)*	.77	.81 (49.6%)

Note. Reference group is students not represented in each category. *r* = zero order correlation. (Unique variance) = squared semi-partial correlation. Tolerance ranged from .83-.99, VIF ranged from 1.01-1.21. * $p < .05$

For the simultaneous LR analyses, student demographic data and the easyCBM® fall score were entered together as independent variables. The passing rates for all students in Oregon were 76.7%, 78.5%, and 69.6% for grades 6-8 respectively. These values were entered as the classification cut off criterion in each analysis. For comparison, the observed percentages of students who met the state standard in our sample are reported in Table 2. In general, the passing rate for our sample is slightly higher compared to the overall passing rate in Oregon across all grades.

Although the independent variables were the same in both analyses, the dependent variables were different. For the multiple regression analysis (MR), students' state test scores were entered as the dependent variable, while students' performance level classification of *does not meet* and *meets or exceeds* was used as the criterion for the logistic regression analysis (LR). There is no direct analogy to an R^2 statistic in LR, and we instead report the results of two tests to evaluate model fit: Deviance and Hosmer and Lemeshow tests. The Deviance test compares the -2Log likelihood value between a null model (without predictors) and the model with all predictors and evaluates the difference of -2Log likelihood value between two models based on the Chi-square distribution. A significant Chi-square value indicates that the predictors have statistical evidence for contributing toward the overall model fit. The Hosmer and Lemeshow test evaluates the alignment between prediction of the model and the actual observation. A non-significant Chi-square value suggests a good model fit. In addition, *Nagelkerke* R^2 is reported as an approximate parallel to R^2 to MR, though it should be interpreted with caution as an approximate indication.

Diagnostic efficiency statistics, including sensitivity, specificity, and positive and negative predictive values, were calculated to provide an indication of the predictive accuracy of the model (Park, Anderson, Irvin, Alonzo, & Tindal, 2011). All statistics were calculated post-hoc from the classification tables obtained through the LR analyses.

Results

Sequential Multiple Regression

Regression coefficients and the unique variance accounted for by each variable are reported by grade level in Table 4. Unique variance was calculated by squaring the semi-partial correlations. Each statistically significant predictor is flagged with an asterisk. Tolerance and VIF are reported in the notes section of Table 4.

Grade 6. The overall regression model for grade 6 was statistically significant. Demographic variables accounted for approximately 16% of the total variance in OAKS scores. The addition of students' fall easyCBM® scores was significant, and increased the proportion of explained variance to 64%. The easyCBM® measure uniquely accounted for 48% of the total variance in OAKS scores and had a significant beta ($b = 1.03$, $p < .05$). On average, for every 1 point students gained on easyCBM®, their corresponding OAKS score increased 1.03 points.

Grade 7. The overall regression model for grade 7

was statistically significant. Demographic variables accounted for approximately 15% of the total variance in OAKS scores. The addition of students' fall easyCBM® scores was significant, and increased the proportion of explained variance to 65%. The easyCBM® measure uniquely accounted for 50% of the total variance in OAKS scores and had a significant beta (b = 0.90, p < .05). On average, for every 1 point students gained on easyCBM®, their corresponding OAKS score increased 0.90 points.

Grade 8. The overall regression model for grade 8 also was statistically significant. Demographic variables accounted for approximately 17% of the total variance in OAKS scores. The addition of students' fall easyCBM® scores was significant, and increased the proportion of explained variance to 66%. The easyCBM® measure uniquely accounted for 49% of the total variance in OAKS scores and had a significant beta (b = 1.02, p < .05). On average, for every 1 point students gained on easyCBM®, their corresponding OAKS score increased 1.02 points.

Table 5. Logistic Regression Coefficients

Grade/Variable	*b*	*SE*	*WALD*	*Odds-Ratio (EXP(B))*
6				
Intercept	-9.04*	0.41	498.52	0.00
Minority	0.20	0.13	2.55	1.23
SPED	0.98*	0.14	47.98	2.66
Female	0.34*	0.12	8.34	1.41
ELL	0.86*	0.20	17.86	2.37
Fall total	0.32*	0.01	514.01	1.38
7				
Intercept	-6.55*	0.36	338.55	0.00
Minority	0.04	0.14	0.09	1.04
SPED	0.81*	0.16	27.39	2.25
Female	0.25	0.13	3.76	1.28
ELL	0.79*	0.24	10.69	2.20
Fall total	0.26*	0.01	436.12	1.30
8				
Intercept	-8.10*	0.37	469.86	0.00
Minority	0.64	0.12	29.00	1.90
SPED	0.89*	0.16	31.89	2.43
Female	0.51*	0.11	20.27	1.67
ELL	0.28	0.24	1.36	1.32
Fall total	0.29*	0.01	577.76	1.34

Note. Reference group is students *not* in each category. * p < .05

Logistic Regression

In LR, the magnitude of the effect of each predictor is often evaluated with an odds ratio. Regression coefficients (in logit scale) and odds ratios for each predictor variable are reported in Table 5. Logistic regression classifications for each grade are reported in Table 6. Diagnostic efficiency statistics are reported in Table 7.

Grade 6. The predictors of this study reflected adequate model fit based on the Deviance test, the Hosmer and Lemeshow test, and Nagelkerke's R^2 (.55). The easyCBM® measure had a significant regression coefficient (b = 0.32, p < .05). On average, a sixth-grade student scoring one point higher than another student would be 1.38 times more likely to pass OAKS. It is important to note that

Table 6. Logistic Regression Classification Results

Observed classification		Predicted Group Membership		Total
		Does not meet	Meets/Exceeds	
Grade 6	Does not meet	567 (82%)	125 (18%)	692 (100%)
	Meets/Exceeds	430 (17%)	2126 (83%)	2556 (100%)
	Total	997 (31%)	2251 (69%)	3248 (100%)
Grade 7	Does not meet	410 (79%)	109 (21%)	519 (100%)
	Meets/Exceeds	368 (15%)	2168 (85%)	2536 (100%)
	Total	778 (25%)	2277 (75%)	3055 (100%)
Grade 8	Does not meet	715 (83%)	151 (17%)	866 (100%)
	Meets/Exceeds	368 (17%)	1850 (83%)	2218 (100%)
	Total	1083 (35%)	2001 (65%)	3084 (100%)

Note. Overall correct classification for grade 6 is 88.4%, grade 7 is 89.0%, and grade 8 is 86.3%.

Table 7. Diagnostic Efficiency of Logistic Regression Classification Results

Grade	Sensitivity	Specificity	PPV	NPV
6	.82	.83	.57	.94
7	.79	.85	.53	.95
8	.83	.83	.66	.92

Note. PPV = Positive Predictive Value; NPV = Negative Predictive Value

when students' scores on easyCBM® increase by more than one point, the odds of passing OAKS increase exponentially. For example, if a student scores 3 points higher on the easyCBM® measure, s/he is 2.63 times more likely to pass OAKS ($1.38^3 = 2.63$).

Grade 7. The predictors of this study reflected adequate model fit based on the Deviance test, the Hosmer and Lemeshow test, and Nagelkerke's R^2(.53). The easyCBM® measure had a significant regression coefficient ($b = 0.26$, $p < .05$). On average, a seventh-grade student scoring one point higher than another student would be 1.30 times more likely to pass OAKS.

Grade 8. The predictors of this study reflected model fit based on the Deviance test, the Hosmer and Lemeshow test, and Nagelkerke's R^2(.59). The easyCBM® measure had a significant regression coefficient ($b = 0.29$, $p < .05$). On average, an eighth-grade student scoring one point higher than another student would be 1.34 times more likely to pass OAKS.

Classification Accuracy and Diagnostic Efficiency

Classification accuracy was compared based on the students' predicted group membership (i.e. *meets* or *exceeds* vs. *does not meet*) based on the logistic regression analysis and their actual performance on the state achievement test. The overall classification accuracy rate was 88.4% for grade 6, 89.0% for grade 7, and 86.3% for grade 8. In contrast, about 15%-17% of students were misclassified as "*not meeting* the standards" when in fact they *met* the standards. Also, about 17%-21% of students were classified as *meeting/exceeding* the standards when they *did not meet* the standards (see Table 6).

Diagnostic efficiency is evaluated by computing sensitivity and specificity as well as positive predictive value and negative predictive value (see Table 7). Sensitivity, also known as true positive rate, ranged from .79 to .82, and specificity, also knowns as true negative rate, ranged from .83 to .85. These results indicate that the diagnostic efficiency of the easyCBM® mathematics measure is satisfactory because, generally, sensitivity and specificity above .70 are considered adequate (Silberglitt & Hintze, 2005).

Discussion

The current study supplements the validity evidence collected to date for using a mathematics CBM within RTI to classify students into instructional tiers early in the school year. Following the logic of Kane (1992, 2006), we empirically examined the assumptions underlying two key inferences for a placement test: extrapolation and decision. In what follows, we review our findings relative to each inference, contrast our findings with previous research, and discuss implications of the robust validity evidence of CBM. We conclude by discussing limitations and directions for future research.

The assumptions underlying the extrapolation inference were investigated by analyzing the relation between students' scores on easyCBM® in the fall (beginning of the school year) and state achievement test scores in the spring (end of the school year). The relation was quite strong. Controlling for demographic variables, the easyCBM® measure uniquely accounted for 48%-50% of the total variance in state test scores. If we accept that state test results are an adequate measure of the annual learning outcomes in schools, we can then infer that easyCBM®, in turn, is quite closely aligned with annual learning outcomes in Oregon, given the strong relation among the measures.

The connection to classroom learning objectives is a critical feature of any CBM because the purpose of CBM is to inform teachers' instructional decision making. If a CBM measures areas irrelevant to students' classroom success, teachers may be misguided into thinking students are at risk when they are actually 'on track.' Similarly, if a CBM does *not* measure relevant areas of classroom success, teachers may be misguided into believing that students are on track when really they are at-risk for future failure.

The assumptions underlying the decision inference for

easyCBM® were investigated by conducting a retrospective analysis. First, we used students' observed performance classification on the state test as the criterion for establishing which students likely needed Tier 2 placement in the beginning of the year. We then used the fall easyCBM® measure to predict students' classification on the state test in the spring – serving as a proxy for tier placement. Results indicated that a student scoring one point higher than another on easyCBM® (standard deviation of easyCBM® ranges from 7.04 to 8.15 across grades 6 to 8) would be 1.30-1.38 times more likely to pass the Oregon state test.

Results from the investigation of the extrapolation inference suggest the fall easyCBM® mathematics measure is predictive of the state test, with the total regression models accounting for over 64% of the total variance in state test scores controlling for student demographic characteristics in each grade.

Despite this strong relation, some students were misclassified, as Table 6 reveals. Approximately 18%-21% of students who *did not meet* expectations were predicted to *meet* or *exceed* (false positive). On the other hand, approximately 15%-17% of students who were predicted to *not meet* the expectations *met* or *exceeded* the expectations (false negative). The overall correct classification rate ranges from 86.3% to 89.0% for grades 6 to 8. It is likely that the false negative rates found were a result of students whose easyCBM® performance placed them in "gray zone" of probability for passing the state test early in the year, perhaps highlighting the importance of using multiple measures to inform classification decisions. False negatives within RTI withhold additional assistance and resources from students who are in need. It is uncertain whether these false negatives are related to CBM in general, easyCBM® in particular, or some other related dimension (e.g. additional resources provided to these students based on the fall CBM scores). It should be noted, however, that the false negative rates observed were those obtained from the LR analyses using the state passing percentage as the probability cutoff level and *not* from a normative cut point within RTI. The observed rates would thus vary substantially based on the RTI model chosen (for optimal cut scores see Anderson, Alonzo, & Tindal, 2010). The classification tables, along with the diagnostic efficiency statistics, do provide an indication of the classification tendencies of easyCBM®.

Previous CBM validity research. The results of this study extend previous CBM validity research in two important ways. First, while other researchers (Espin et al., 2009; Keller-Margulis et al., 2008; Shapiro et al., 2006) have examined the relation of a mathematics CBM used early in the year with statewide test performance later in the year, the sample sizes were quite small. In contrast, the sample size in the current investigation was large, ranging by grade from 3,084 – 3,248. Although the large sample does not guarantee the generalizability, the computed statistics are likely more stable than in prior studies conducted with smaller samples.

Espin et al. (2009), Shapiro et al. (2006), and Keller-Margulis et al. (2008) all reported moderately strong correlations between CBMs administered early in the school year and statewide test performance later in the school year. Espin and others reported coefficients ranging from .44 to .55; Shapiro and colleagues reported coefficients ranging from .07 to .64; and Keller-Margulis and associates reported correlations ranging from .14 to .66. In the current study, the correlation coefficients were larger than those found by any of these studies, ranging from .79 to .82. The observed correlations were near the low end of those typically observed in CBM alternate form studies (see Foegen et al., 2007).

The high correlations found may reflect the process used in easyCBM® item development, during which items were specifically written to align with the NCTM grade-level focal points. This process marks a substantial shift in CBM development, moving away from the more common mathematics CBMs assessing skills in computation or the application of concepts to test development that borrows heavily from approaches used in statewide mathematics test development. Given that (a) most states use curricular standards when constructing the state achievement tests and (b) instruction is (presumably) focused around the standards, CBMs developed to be aligned with curricular standards may more accurately represent the content of classroom instruction, providing further evidence for the assumptions underlying the extrapolation inference.

Second, this study used two statistical analyses unique from previous research to better understand how the CBM measure relates to the state test while controlling for the impact of other potential sources of systematic error. Using multiple linear regression and logistic regression at each grade level, we were able to provide more in-depth analyses and document the unique contribution of each

predictor variable (linear regression) and important information on the measure's discriminating power between students who *met* and *did not meet* the standard on the state test (logistic regression). In addition, the analyses allowed us to control for student demographic variables, a potential source of systematic error.

Using regression analyses, the validity of decisions and inferences based on the measures can be drawn more clearly. In the current study, for instance, the semi-partial correlations (the correlation between easyCBM® and state test performance while controlling for demographic variables) ranged from .69-.71. These correlations are slightly lower than the zero-order correlations, but the validity argument remains strong because the unique contribution of the measures is still quite high.

Implications of robust validity evidence of CBM for improving practice. Accurate screening of students for proper instructional tier placement is one of the most important factors for successful implementation of RTI. Use of CBM assessments with robust validity evidence is certainly one way to increase the accuracy of this process. When CBM is used in practice, a careful examination of content and constructs measured by CBM is necessary to make sure the test does not measure skills irrelevant to classroom instruction. Also, the alignment of content and constructs assessed in CBM and instructional focus and curriculum standards is a requisite. Without such alignment, assessment data collected from CBM screeners may not provide valid and useful information for teachers to improve instructional practices.

Use of CBM within the RTI framework always carries the possibility of student misclassification no matter how great the measure is. That is because the misclassification of student is not just a product of a particular CBM assessment, but it is also a function of the cut score used for classification. Therefore, it is important to carefully evaluate the availability of school resources and the capacity of instructional staff when adopting a specific cut score. Given that a 100% accurate classification is an unrealistic expectation, schools need to balance the risk of wasting resources by over-classifying (i.e. setting the cut score high such as at the 40th percentile) versus failing to provide instructional support to students who are in need by under-classifying (i.e. setting the cut score low such as at the 10th percentile).

Limitations. Findings from this study should be interpreted relative to a few important limitations. The limitations are related to the variables that were studied and the location of the study. First, tier placements, and any subsequent instructional effects, were not accounted for in these analyses. Because we did not account for instruction or tier placement, the results are inherently exploratory. However, it is also logical to assume that accounting for an instructional effect would only strengthen the overall predictive power of the model. Had we included a variable such as tier placement or instructional support provided to students, it is likely that fewer students who passed the state test would have been predicted to not pass.

Second, the analyses reported here investigated the relation between easyCBM® and the state assessment in Oregon. Thus, while the sample was relatively large, it was drawn exclusively from the Pacific Northwest and may not generalize to other populations as the obtained sample differs systematically from other districts, states, and educational contexts. Needless to say, these concerns should be addressed in future studies to supplement the results of this investigation, which would further extend the validity evidence related to using easyCBM® within an RTI framework. However, it is important to keep in mind that the validity argument of CBM does not lie with a particular cut point for a particular assessment or curriculum, but it rather emphasizes the importance of valid use of CBM as a classification tool to improve accuracy of instructional placement and enhance allocation of school resources.

Conclusions and directions for future research. The overall evidence strongly supports the use of the fall easyCBM® mathematics measure as a classification tool within RTI for grades 6-8. State test performance, however, should not be the only criterion used in establishing tier placement, even though it was the only criterion used in the current investigation. Rather, it is important for educators to consider other factors in their decision-making, such as previous documentation of student work, slope of improvement, and teacher judgments. Future studies should continue to explore the validity of CBM within RTI. In particular, continued examination of the consistency and accuracy of CBM classifications within RTI appears to be an area of need.

Future studies should control for other sources of variance in an attempt to explain these misclassifications. Controlling for tier placement, instructional effects, and student growth over time may increase the accuracy of

classification. Eventually, probabilities for proficiency may be embedded within a computer-based CBM system to give teachers an indication of the likelihood that each student in their class would pass the state test if no intervention were provided. However, before this sort of information could be provided to teachers, the classification accuracies need to be considerably improved, an area of research in its infancy.

References

Anderson, D., Alonzo, J., & Tindal, G. (2010). *Diagnostic efficiency of easyCBM: Oregon State* (Technical Report #1009). Eugene, OR: Behavioral Research and Teaching.

Burke, M., Hagan-Burke, S., Kwok, O., & Parker, R. (2008). Predictive validity of early literacy indicators from the middle of kindergarten to second grade. *Journal of Special Education*, 42(4), 209-226. doi: 10.1177/0022466907313347

Chard, D., Clarke, B., Baker, S., Otterstedt, J., Braun, D., & Katz, R. (2005). Using measures of number sense to screen for difficulties in mathematics: Preliminary findings. *Assessment for Effective Intervention*, 30(2), 3-14. doi: 10.1177/073724770503000202

Christ, T., Johnson-Gros, K., & Hintze, J. (2005). An examination of alternate assessment durations when assessing multiple-skill computational fluency: The generalizability and dependability of curriculum-based outcomes within the context of educational decisions. *Psychology in the Schools*, 42(6), 615-622. doi: 10.1002/pits.20107

Christ, T., & Vining, O. (2006). Curriculum-based measurement procedures to develop multiple-skill mathematics computation probes: Evaluation of random and stratified stimulus-set arrangements. *School Psychology Review*, 35(3), 387-400.

Deno, S. (2003). Developments in curriculum-based measurement. *The Journal of Special Education*, 37(3), 184-192. doi: 10.1177/00224669030370030801

Espin, C. A., Wallace, T., Foegen, A., Du, X., Ticha, R., Wayman, M. M., et al. (2009). *Seamless and flexible progress monitoring: Age and skill level extensions in math, basic facts* (Technical Report #2): Research Institute on Progress Monitoring, University of Minnesota.

Fewster, S., & Macmillan, P. (2002). School-based evidence for the validity of curriculum-based measurement of reading and writing. *Remedial and Special Education*, 23(3), 149-156. doi: DOI: 10.1177/07419325020230030301

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *Journal of Special Education*, 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, L., & Fuchs, D. (2007). A model for implementing responsiveness to intervention. *Teaching Exceptional Children*, 39(5), 14-20.

Fuchs, L., Fuchs, D., Compton, D., Bryant, J., Hamlett, C., & Seethaler, P. (2007). Mathematics screening and progress monitoring at first grade: Implications for responsiveness to intervention. *Exceptional Children*, 73(3), 311-220.

Goffreda, C., Diperna, J., & Pedersen, J. (2009). Preventive screening for early readers: Predictive validity of the dynamic indicators of basic early literacy skills (DIBELS). *Psychology in the Schools*, 46(6), 539-552. doi: 10.1002/pits.20396

Helwig, R., Anderson, L., & Tindal, G. (2002). Using a concept-grounded, curriculum-based measure in mathematics to predict statewide test scores for middle school students with LD. *Journal of Special Education*, 36(2), 102-112. doi: 10.1177/00224669020360020501

Hintze, J., Christ, T., & Keller, L. (2002). The generalizability of cbm survey-level mathematics assessments: Just how many samples do we need? *School Psychology Review*, 31(4), 514-528.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527-535. doi: 10.1037/0033-2909.112.3.527

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (Fourth ed., pp. 17-64).

Keller-Margulis, M. A., Shapiro, E., & Hintze, J. (2008). Long-term diagnostic accuracy of curriculum-based measures in reading and mathematics. *School Psychology Review*, 37(3), 374-390.

Nese, J. F. T., Lai, C. F., Anderson, D., Jamgochian, E. M., Kamata, A., Sáez, L., et al. (2010). *Technical adequacy of the easyCBM mathematics measures, (Grades 3-8), 2009-2010 version* (Technical Report #1007). Eugene, OR: Behavioral Research and Teaching.

Nese, J. F. T., Lai, C. F., Anderson, D., Park, B. J., Tindal, G., & Alonzo, J. (2010). *The alignment of easyCBM math measures to curriculum standards* (Technical Report No. 1002). Eugene, OR: Behavioral

Research and Teaching.

Oregon Department of Education. (2006-2007). *Technical report: Oregon's statewide assessment system: Reliability and validity* (Vol. 4). Salem, OR.

Oregon Department of Education. (2008). *Oregon's new core standards structure: Mathematics leads the way*. Salem, OR.

Oregon Department of Education. (2010). Assessment scoring: Frequently asked questions about scoring statewide assessments Retrieved April 21, 2010, from http://www.ode.state.or.us/apps/faqs/index.aspx?=88

Park, B. J., Anderson, D., Irvin, P. S., Alonzo, J., & Tindal, G. (2011). *Diagnostic efficiency of easyCBM reading: Oregon* (Techincal Report No. 1106). Eugene, OR: Behaviroal Research and Teaching.

Shapiro, E., Keller, M., Santoro, L., & Hintze, J. (2006). Curriculum-based measures and performance on state assessment and standardized tests. *Journal of Psychoeducational Assessment*, 24, 19-35.

Silberglitt, B., & Hintze, J. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment*, 23(4), 304-325. doi: 10.1177/073428290502300402