The Version of Record of this manuscript has been published and is available in *School Psychology Review*, 2022, http://www.tandfonline.com/10.1080/2372966X.2022.2041211

Predicting Interim Assessment Outcomes Among Elementary-Aged English Learners Using Mathematics Computation, Oral Reading Fluency, and English Proficiency Levels

Garret J. Hall[a], Mitchell A. Markham[b], Meghan McMackin[c], Elizabeth C. Moore[b], & Craig A. Albers[b]

[a]Florida State University

[b]University of Wisconsin-Madison

[c]Austin Anxiety and OCD Specialists

Author Note

M. A. M., M. M., E. C. M., and C. A. A. share second authorship. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Awards #R305B150003 to the University of Wisconsin- Madison and # R305A100585 to Craig A. Albers. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education.

M.A.M. is now at Madison Metropolitan School District, Madison, WI, and E. C. M. is now at Nationwide Children's Hospital, Columbus, OH. Correspondence concerning this article should be addressed to Garret J. Hall, Department of Educational Psychology and Learning Systems, Florida State University, 1114 W. Call St, 3204H Stone Building, Email: gjhall@fsu.edu

Declaration of interests: none

Abstract

The current study examined the validity of curriculum-based measures (CBM) in mathematics computation (M-COMP) and oral reading fluency (R-CBM) in predicting spring mathematics and reading performance level and performance risk (>1 *SD* below the national mean) among students classified as English Learners (ELs). Additionally, the current study assessed the incremental predictive value of English language proficiency (ELP) beyond CBM performance. The results indicated that ELP explains a significant portion of variability above M-COMP and R-CBM and increases the accuracy of predicting at-risk performance status on spring measures of mathematics and reading. The findings highlight the challenges of assessing the predictive accuracy of M-COMP and R-CBM among students classified as ELs, as well as the extent to which comprehensive measures of ELP account for variance in both performance level and at-risk status beyond CBMs. The implications for school data-based decision-making for language-minoritized students and directions for future research are discussed.

Keywords: mathematics computation, oral reading fluency, English language learners, English language proficiency

Impact statement:

Equity in Response-to-Intervention (RTI) is predicated on accurate measurement of skills within universal screening. The current study's findings suggest that CBMs alone explain less variance and are less predictive of academic performance than when combined with English language proficiency scores. The predictive accuracy of R-CBM and M-COMP varied between students classified as ELs and non-ELs but in only very limited circumstances were these measurable differences. These results indicated that although CBMs are an efficient system of screening

among non-ELs, it is also necessary to consider students' ELP levels when making decisions within RTI models.

Predicting Interim Assessment Outcomes among Elementary-Aged English Learners using Mathematics Computation, Oral Reading Fluency, and English Proficiency Levels

Response-to-Intervention (RTI) is a multitiered educational service delivery model that aims to identify students at risk for academic difficulties and provide these students with supplemental evidence-based interventions (Fletcher & Vaughn, 2009). RTI is empirically supported (Burns et al., 2005; Fien et al., 2021) and has been widely promoted in both policy (e.g., Every Student Succeeds Act, 2015-2016; Individuals with Disabilities Education Improvement Act [IDEA], 2004) and practice (e.g., the National Association of School Psychology's [NASP] Practice Model; NASP, 2020). Because RTI involves a variety of techniques and ideas, it manifests differently across schools and districts. However, common elements of high-quality RTI models include (a) evidence-based core classroom instruction, (b) data-based decision procedures, (c) universal screening procedures, (d) progress monitoring procedures, (e) clearly defined processes for movement between service delivery tiers, (f) evidence-based supplemental interventions, (g) intervention integrity procedures, and (h) clear procedures for determining special education eligibility as part of a comprehensive assessment (Albers & Martinez, 2015).

A primary goal of RTI models is to support the early identification of all students who are experiencing academic challenges. Therefore, all students should be able to equally access universal and tiered supports such that prevention and early intervention also are equitably applied with students from language-minoritized backgrounds. Students from language-minoritized backgrounds are often identified by state and local educational authorities as English learners (ELs) and are a student population subgroup that may be inequitably served in many

ways in schools (Albers & Martinez, 2015; Robinson-Cimpian et al., 2016), which therefore may contribute to the poorer academic achievement often referenced among this student population.

EL Definitions and Educational Characteristics

As defined by federal and state statutes, ELs are defined as primary or secondary school students whose native language is not English (Every Student Succeeds Act, 2015-2016), and who are therefore proficient (or reaching proficiency) in other languages and also have emerging English-language proficiency.[1] Demographic data have indicated that the school-aged EL population is one of the most rapidly growing subgroups in schools throughout the United States, with 10.1% of school-aged students being classified as ELs in 2017 (Hussar et al., 2020).

Research has also suggested that academic outcomes for this growing student population tend to be, on average, lower than academic outcomes associated with non-EL students. For example, ELs have been reported to achieve lower academic test scores than their non-EL peers (Hussar et al., 2020), and this trend among ELs is similar trend among language minoritized students (Roberts et al., 2010; Roberts & Bryant, 2011). This has various implications associated with special education placement and resulting overrepresentation (i.e., disproportionality) of ELs within special education. However, disproportionately data in special education appear highly dependent on the definition used to capture linguistic diversity (e.g., EL compared to broader definitions of linguistic diversity; Morgan et al., 2018; Samson & Lesaux, 2009; Sullivan, 2011; Sullivan & Bal, 2013; Umansky et al., 2017). For example, when Kieffer and Thompson (2018) used a categorization of "multilingual" to examine Grade 4 and Grade 8 NAEP trends over time, they found greater improvements in NAEP scores among multilingual

[1] The category of EL comprises only a subset of students who could be multilingual or have differing levels of proficiency in multiple languages. When describing studies that focus on linguistically diverse students (not limited to ELs), we use the term "language minoritized", unless the sample in those studies consists only of students who meet the federal definition of EL (in which case we use the term EL).

students relative to English-monolingual students. These conclusions differed substantially as compared to using federal and state EL definitions. Consequently, the research on disproportionality in special education placement, combined with the academic achievement gaps between ELs and non-ELs, suggests that greater precision is needed in identifying the risk and protective factors that may impede or support these ELs' educational successes in order to more adequately identify and intervene when the need for additional academic supports is indicated. In theory, one way to accomplish this is using reliable, valid, and equitable universal screening approaches that take into consideration current levels of English language proficiency.

Universal Screening

From an RTI perspective, one of the first and most efficient steps in enhancing this precision in data-based decision making for academic achievement is universal screening (Albers & Martinez, 2015; Glover & Albers, 2007). Academic universal screening is a widely used assessment method that takes on a variety of formats, but typically entails the assessment of a skill that serves as an indicator of development towards proficiency in a particular academic area (Hosp et al., 2016). Universal screeners are designed to identify students with performance or skill deficits (specifically in areas of automaticity or fluency). This early identification is then intended to lead to additional supplemental instructional supports, so it is important to consider the effectiveness and efficiency of screeners, as well as their relationship to intervention (Clemens et al., 2016; Glover & Albers, 2007). If a screener has adequate predictive validity – which is the capacity of a screener to accurately predict risk status or explain variance on an outcome measured at a later time – this predictive capacity should help practitioners identify appropriate tiered supports and corresponding intervention dosage.

Curriculum-based measures (CBMs) are widely used tools that provide valuable academic assessment data. These measures are commonly utilized in the universal screening of students' academic skills given their simplicity, efficient use, easy administration, and interpretability (Deno, 2003). Two CBMs that are often used within schools include Reading-CBMs (R-CBMs) and Math-CBM (M-CBMs). The use of R-CBMs to predict overall reading achievement has been consistently supported by research (e.g., Keller-Margulis et al., 2008; Van Norman & Nelson, 2021), with one of the most frequently used R-CBMs being oral reading fluency (ORF). M-CBMs measure a range of skills, including basic number identification and discrimination, basic and complex computational skills, and the application of mathematical concepts. Although M-CBMs are frequently used within school settings, they have received significantly less research attention and empirical support than R-CBMs. As one form of M-CBM, Mathematics Computation (M-COMP; Pearson Education & Inc, 2012) has been identified as being technically adequate (Christ et al., 2008; Thurber et al., 2002) and effective for use with non-EL populations (Keller-Margulis et al., 2008; Vanderheyden & Burns, 2005). Thus, although the use of R-CBM (including ORF) and M-CBM (including M-COMP) as universal screeners of academic skills is empirically supported for non-EL populations, less is known about their use with students from language-minoritized backgrounds.

To accurately identify all students with or at risk for academic skill deficits, including students classified as ELs, the reliability and validity of reading, math, and other CBMs must be adequately examined. Early evidence suggested that R-CBMs were similarly reliable for ELs as their non-EL peers (Baker & Good, 1995) and were predictive of their performance on high-stakes testing (Kim et al., 2016; Muyskens et al., 2009; Wiley & Deno, 2005). Considering the policy requirements of the Every Student Succeeds Act (2015-2016) to track growth toward

English language proficiency (ELP) among students classified as ELs, an underexplored area of examining predictive validity is the incremental contribution of ELP beyond ORF or M-COMP in predicting student achievement and risk status. Although the complexity of ELP may differ substantially from language measures used in other studies (e.g., phonological skills, syntax, receptive/expressive vocabulary), the practical imperative to monitor ELP development necessitates greater attention to ELP's role within universal screening. Given the focus of universal screening on efficiency, cost-effectiveness, and equity (Albers & Martinez, 2015), building greater understanding of the role(s) that ELP plays in relation to ORF, and especially M-COMP considering the limited research base, is necessary. Below, we provide additional background regarding the skills necessary for reading and mathematics achievement, their relationship to ELP, and how M-COMP and ORF may help identify deficits within screening.

ORF is a widely-used screening tool that taps a number of underlying reading skills. Evidence for the underlying component skills of ORF, as a measure of text reading fluency, and ORF's relation to comprehension has grown in recent years (e.g., Kim, 2012; Kim & Wagner, 2015). Generally, ORF is thought to link word reading and language comprehension skills to reading comprehension (Kim & Wagner, 2015). ORF recruits individuals' capacity to quickly orally decode words while simultaneously applying acquired vocabulary knowledge to construe meaning from the text (Kim & Wagner, 2015). However, these underlying patterns might not similarly hold among ELs, requiring further investigation into how ELs' ELP and ORF performance relate to reading performance. Kim (2012) tested a model among elementary Spanish-speaking ELs wherein English word reading was directly related to ORF, and English oral language as well as oral reading fluency were both related to English reading comprehension. Kim's (2012) findings were inconsistent with the author's prior work, which

showed that word reading and oral language predicted text reading fluency (Kim et al., 2011). The general conclusion from Kim (2012) was that ORF might recruit word reading and language skills differently for ELs than English-proficient students, bringing into question what other variables might relate to ELs' English reading skills.

Prior to Kim (2012), Lesaux et al. (2010) and Gottardo and Mueller (2009) both found that English oral language predicted English reading comprehension over and above Spanish word reading and Spanish oral language, although Gottardo and Mueller found that English word reading also predicted English comprehension, whereas Lesaux et al. (2010) did not find this relationship. Among other important methodological differences between the studies (e.g., Gottardo and Mueller's sample include first and second grade students; Lesaux et al.'s sample included fifth grade students), Lesaux et al. included a measure of word reading fluency in their English and Spanish word reading latent variables, whereas Gottardo and Mueller did not. Given the role of automaticity in reading development, this might be a potentially meaningful difference in the models between Lesaux et al. and Gottardo and Mueller. Another important recent finding indicates that academic language growth in Grades 6–7 related to reading comprehension growth across the same time period controlling for EL status as well as other demographic characteristics (Galloway & Uccelli, 2019). The same study also found that growth rates of English academic language among ELs (whom the authors defined as emergent bilinguals) and English-proficient students through Grades 6–7 were similar, although ELs performed significantly below their English-proficient peers throughout this time period. Altogether, these results underscore the importance of considering the unique relationships of ORF and language to reading skills among ELs.

Generally, ORF has received more attention than M-COMP with respect to validity among students classified as ELs (e.g., Kim et al., 2016; Vanderwood et al., 2014). Research examining the reliability and validity of M-CBMs, specifically M-COMP, with EL or language-minoritized populations is sparse and requires additional attention given the unique relationship between language and mathematical skills. In particular, symbolic mathematical knowledge has been posited to emerge in part from language-based processes (Dehaene, 1992; LeFevre et al., 2010; Vukovic & Lesaux, 2013a). In a recent meta-analysis, Peng et al. (2020) found a moderate correlation (r = .42) between mathematics and language across 344 studies, suggesting that language and mathematics are meaningfully correlated, requiring direct attention to this relationship when considering screening for mathematics (i.e., mathematics skills cannot be considered in isolation).

Supplementing these findings, multiple studies have focused on specific language subskills and their relations to mathematics outcomes. Among Spanish-English dual language learners in preschool, Méndez et al. (2019) showed that expressive vocabulary and grammar comprehension robustly predicted early numeracy performance within each language. Among second and third grade students, English receptive syntax abilities, morphology, and a language screening tool capturing multiple domains predicted mathematics achievement (Chow & Ekholm, 2019). Other research has suggested that English oral language ability in elementary school facilitates performance on conceptually-focused (but not arithmetically-focused) mathematics tasks among both native English-speaking students and language minoritized students (Vukovic & Lesaux, 2013b). Vukovic and Lesaux (2013a), although not focused specifically on linguistic diversity, found that third graders' verbal skills were related to arithmetic skills through symbolic number knowledge, and phonological skills related directly to

arithmetic skills. Consistent with theory (Dehaene, 1992) and prior empirical models (LeFevre et al., 2010), Vukovic and Lesaux's (2013a) findings suggested that even basic language skills provide a foundation for certain aspects of mathematical cognition, though some of these relations may be indirect. As previously noted, Peng et al.'s (2020) findings provided evidence for the connections between language and mathematics. However, as the other studies mentioned suggest, different language domains may differentially relate to a variety of mathematics skills. On balance, these studies' findings indicate a need to consider the multiple sources of mathematical reasoning skills (e.g., language as well as content-specific mathematics knowledge) and how these variables factor into data-based decision making in universal screening, particularly when screening students with a variety of language backgrounds and proficiencies.

Connecting ELP to Decision-Making in RTI

RTI's foundational concepts of prevention and early intervention provide promise for its ability to improve the outcomes of students classified as ELs (Allsopp et al., 2016). However, the complexities of the ELs' experiences in school pose special challenges within each component of RTI, which is why Albers and Martinez (2015) proposed the consideration of ELP level and language acquisition factors (e.g., basic interpersonal communicative skills, cognitive academic language, academic language, classroom language instructional model, mobility rates) at each stage within the RTI process. Current measures of ELP are designed to assess the use of social and academic language across various components of language use (e.g., reading, writing, listening, speaking) and their administration with ELs is required to meet federal Title I and Title III accountability requirements for monitoring student progress. Although a small number of states use either the English Language Proficiency Assessment for the 21st century (ELPA21;

CRESST, n.d.) or state-specific ELP measures (Albers & Martinez, 2015), the most commonly used ELP measure is the ACCESS for ELLs, which is currently used in 40 states and territories (WIDA, 2020). ELP measures have been shown to have varying levels of focus on social language versus academic language, content focus (e.g., mathematics, science), and correlation to performance in specific content areas (Wolf & Faulkner-Bond, 2016). This variance between measures creates challenges in the interpretation of results and subsequent decision-making. Consequently, there are challenges inherent to the assessment of language-minoritized students, including students classified as ELs according to federal definitions, for risk identification (O'Bryon & Rogers, 2010) because it is often difficult to disentangle the impact of language acquisition from other factors, such as inadequate instruction or the presence of disorders and disabilities (Harry et al., 2014**)**. Accurate assessment processes are vital to support proper risk identification within the RTI model.

Given the role of language in academic skill development (Galloway & Uccelli, 2019), as well as the need to curb achievement gaps between students classified as ELs and their native English-speaking peers, it is important to identify the extent to which existing universal screening procedures adequately predict the need for intervention and align with other existing measures of assessment and accountability. A number of forms of technical adequacy for screeners exist (Glover & Albers, 2007), although for the purposes of screening, predictive validity is especially meaningful. Sensitivity and specificity are core measures of predictive adequacy. *Sensitivity* refers to the ability of a measure to correctly identify individuals who are at risk, whereas *specificity* refers to individuals who are not at risk (Glover & Albers, 2007); thus, tools that effectively distinguish those truly at risk (i.e., sensitivity) and those not truly at risk (i.e., specificity) are able to provide practitioners with optimal information regarding students

who warrant follow-up assessment or intervention in the skill area assessed. A better understanding of the strengths and limitations of current screening systems will lend itself to earlier risk identification, and thus, more targeted skill development among ELs.

Current Study

The purpose of the current study was to examine the predictive validity of M-COMP and ORF for elementary-aged students classified as ELs and the incremental contribution of ELP in predicting later performance level and risk. Although CBMs are frequently used for instructional and placement decisions for ELs, ELP measures are less directly tied to RTI given the different policy statutes driving the use of each assessment, despite their purpose for informing instruction and intervention. Additional research is needed to understand the relationships between CBMs and ELP data to inform data corroboration, as ELP has potentially substantial benefits in understanding how to target additional assessment and intervention provision within RTI (Albers & Martinez, 2015). The current study investigated the predictive validity of fall academic screeners (i.e., AIMSweb M-COMP and ORF) and winter ELP data (i.e., ACCESS) for performance on a spring interim academic assessment (i.e., Measures of Academic Progress [MAP]) among students classified as ELs in Grades 1–5.

Specifically, we examined the following research questions to better understand the relationships described above:

1. What is the incremental value of including ELP scores with M-COMP and R-CBM in Grades 1–5?
2. Are M-COMP and R-CBM screening measures less predictive among students classified as ELs as compared to non-ELs, particularly after accounting for ELP among ELs?

We hypothesized that ELP data would explain more variance and increase predictive accuracy more so than CBMs alone, suggesting that ELP holds a unique predictive capability beyond CBMs. We predicted that CBMs would not be as predictive of academic performance among ELs as compared to non-ELs, consistent with prior work (e.g., Vanderwood et al., 2014).

Method

Participants

Schools in two Midwestern and one southern state in the United States participated in the current study. The current data were drawn from a larger study of students in kindergarten through twelfth grade. Data were drawn from the 2010–11, 2011–12, and 2012–13 school years. We use data from ELs as well as non-ELs. ELs are those students who met the federal definition of EL: they spoke a language other than English at home and did not meet the state-specified criteria for English proficiency based on their ELP assessment score. We do not have data for students' socioeconomic status or primary home language.

The sample of ELs (Grade 1 n = 143, Grade 2 n = 125, Grade 3 n = 111, Grade 4 n = 71, Grade 5 n = 77) was approximately 50% female across grades, with between 3% (Grade 1) and 25% (Grade 5) having ever received special education services. The sample of ELs was between 49% (Grade 1) and 62% (Grade 4) Black/African American, between 14% (Grade 5) and 20% (Grade 1) Asian, and between 4% (Grade 3) and 9% (Grades 1 and 5) Hispanic/Latino/a. Remaining percentages of students were Native American, Pacific Islander, or White.

We also used data from non-ELs, who were students who did not meet the federal definition of EL (Grade 1 n = 1,202, Grade 2 n = 1,539, Grade 3 n = 1,465, Grade 4 n = 1,424, Grade 5 n = 1,415). In Grades 1–5, between 11%–15% of non-ELs received special education services and 48%–52% were female. Approximately 2%–4% of students were Black/African

American, 1% were Asian, 2% were Hispanic/Latino/a, 2% were Native American, and 91%–93% were White. A small number of individuals were classified as Pacific Islander. Non-ELs scored around the 50th percentile on R-CBM and M-COMP in the fall of Grades 2–5 and in winter of Grade 1. Across all grades, non-ELs scored approximately one-third of a standard deviation above the spring MAP Mathematics and Reading averages. In Grades 1–3, students attended eight different schools; in Grades 4–5, students were drawn from nine schools.

Measures

ELP Level

The ACCESS for ELLs (hereafter referred to as ACCESS) was administered by each participating school district each winter to determine students' ELP levels. The ACCESS is a standardized, summative assessment that collects English language data for each EL student in speaking, listening, reading, and writing domains (Center for Applied Linguistics, 2015). The ACCESS combines each students' performance in the four domains (Reading, Writing, Listening, Speaking) to produce the Overall Composite score. The ACCESS is a vertically scaled measure designed to measure growth from kindergarten through twelfth grade across a continuum of scores that ranges from 100 to 600 (Center for Applied Linguistics, 2015). Within grade levels, the Overall Composite score is converted into proficiency scores (range: 1.0–6.0) that correspond to qualitative indicators of a student's ELP development: (1) entering, (2) beginning, (3) developing, (4) expanding, (5) bridging, and (6) reaching (Gottlieb et al., 2007). ACCESS Overall Composite score reliability was high across applicable grade bands, including .94 for Grades 1–2 and .93 for Grades 3–5 during the 2012–2013 academic year (Center for Applied Linguistics, 2014). Earlier convergent validity estimates (Albers et al., 2009) for the ACCESS ranged from .54 to .74 for the IDEA Proficiency Test (Ballard & Tighe Publishers,

1991), .32–.76 for the Language Assessment System (Data Recognition Corporation, n.d.), .53–.82 for the Language Proficiency Test Series (MetriTech, Inc., n.d.), and .18–.84 for the Maculaitis Assessment of Competencies Test of English Language Proficiency (Maculaitis, n.d.). The current study used the ACCESS composite scale score to assess the incremental contribution of ELP in predicting achievement outcomes. Proficiency levels (i.e., the transformed ACCESS Overall Composite scale score that ranges from 1.0 to 6.0, rounded to the nearest tenth) may be more practically relevant for data-based decision-making considering the proficiency levels typically form the proficiency criteria. However, the scale score offers a more granular measure of ELP. Only a subset of the students in our data had ACCESS scale scores and thus were included in the analyses; other students had ELP levels available, but not scale scores, because a number of schools only provided scale scores in specific years.

Interim Assessment of Academic Achievement

The Measures of Academic Progress (MAP; Northwest Evaluation Association [NWEA], 2011) is a computer-adaptive reading (i.e., word meaning and vocabulary knowledge, understanding and integrating key ideas and details, understanding and interpreting craft and structure) and mathematics (i.e., operations and algebraic thinking, numbers and operations, measurement and data, geometry) interim assessment administered to students in Grades K–12 to assess academic progress along a vertical Rash unit (RIT) scale. MAP test-retest reliability estimates have ranged from .70 to .87, internal consistency from .67 to .93, construct validity from .40 to .57, and concurrent validity from .66 to .88 with the sample reflecting the general student population within the United States (NWEA, 2004, 2011). For the purposes of the current study, we defined academic risk using MAP by identifying students scoring > 1 *SD* below the national normative mean in Spring of each grade based on 2015 norms (i.e., we

subtracted the Spring *SDs* from the Spring normative means for reading or math in Grades 1–5 reported in NWEA, 2015), rounded the difference to the nearest whole number, and assigned a 1 to any student scoring below the calculated value [0 otherwise]). We chose this criterion because 1 *SD* below the national norm represents a score that falls below the typical definition of the national normative average range (i.e., +/- 1 *SD* of the national mean), and this criterion strikes a balance between more or less restrictive definitions of risk or difficulty found in the literature (e.g., Fuchs et al., 2021; Geary, 2010; Gersten et al., 2005). However, given the range of definitions of risk or difficulty used (Nelson & Powell, 2018), our criterion is nonetheless a limited representation of risk status.

R-CBM

AIMSweb R-CBM (Pearson Education & Inc, 2012) is a brief benchmark assessment of ORF and is measured by words correct per minute (WCPM). WCPM calculations are well-established in the literature as having strong psychometric properties (January et al., 2018; Poncy et al., 2005). AIMSweb R-CBM is a norm-referenced and standardized tool that schools may use for both universal screening and progress monitoring purposes. The AIMSweb technical manual (Pearson Education & Inc, 2012) reported AIMSweb R-CBM inter-rater reliability of .99, split-half reliability between .93–.95, and criterion validity between .47–.81, indicating that AIMSweb R-CBM is a reliable and valid measure of oral reading fluency for the general student population. Throughout the results described below, R-CBM is referred to as ORF to emphasize the specific skill domain. As per the standardized administration protocol, all students received three probes of the ORF assessment, which each probe lasting 1 min, at each triannual assessment wave in English. To be consistent with AIMSweb administration guidelines, we used the median score from the three administered probes in all analyses.

M-COMP

AIMSweb M-COMP (Pearson, 2012) is a brief benchmark assessment of mathematics computation that assesses students' fluency with basic computation facts. Students are allotted 8 min to complete the paper-based, open-ended computation problems, which were administered in English. Cronbach's coefficients in Grades 1–5 were reported as ranging from .82 to .91, with split-half reliabilities reported as ranging between .85–.93 (Pearson, 2012). Additionally, Pearson (2012) reports M-COMP's correlation with a summative, criterion-referenced measure of mathematics performance (Group Mathematics Assessment and Diagnostic Evaluation) was .84 in Grade 1 and .73 in Grade 3. These findings suggest that M-COMP is a reliable and valid measure of mathematics computation fluency.

Procedure

Data Collection

All procedures were approved by schools and the primary university institutional review committee. The data presented here came from a larger data collection effort in which a research team partnered with school districts in three states over four years (2010–2011 to 2013–2014) to collect triannual universal screening CBM data in reading, mathematics, and writing, in addition to English language proficiency data and other academic assessment data as available (e.g., MAP, state assessment data). School districts were located in rural, suburban, or urban areas. Analyses presented here include data from 2010–11 to 2012–13 (adequate data were not present for 2013–14 to include this year in the analysis). Some data were already collected by schools and were shared with university researchers whereas other data were collected within schools by members of the research team. In most schools, universal screening CBM data were collected in the fall, winter, and spring. M-COMP and R-CBM were administered starting in the winter of

Grade 1 and were administered in the fall in Grades 2–5. Other interim assessment data (i.e., MAP) that schools collected as part of standard practice were typically collected in the fall and spring. Spring MAP scores were used as the outcome measures in the current study. The ACCESS measure was administered in the winter of each school year.[2]

Analysis Plan

We first conducted ordinary least squares (OLS) regression analyses to assess the linear predictive capacity of CBM measures for MAP outcomes, specifically attending to R^2 values. Then, we assessed the incremental contribution of ACCESS scores to the OLS models by examining adjusted R^2 values and testing the improved model fit with omnibus F-tests of differences in the models.

The last set of analyses consisted of logistic regressions in which we used the same procedure for testing these models as we used to test the OLS models. Using the predicted probabilities of classification as at risk on Spring MAP measures, we then constructed ROC curves to evaluate each model's ability to accurately classify students' performance status. All ROC curves were constructed and assessed using the *pROC* package (Robin et al., 2011) in R (R Core Team, 2021). Specifically, the *ggroc* function of *pROC* and *ggplot2* (Wickham, 2016) were used to construct ROC curves from the predicted probabilities generated from the logistic regressions. Students with a predicted probability of greater than .50 were classified as at risk. Although there are limitations to this criterion, we use it as a neutral indicator of what constitutes

[2] It is important to note that ACCESS scores are administered in the winter and scores are typically reported back to schools in late spring. As a result, ACCESS scores from the prior year typically inform ELL status in any given current year. However, students were not followed consistently across school years in the current study, so we use ACCESS scores from the student's current year (i.e., Grade 1 ACCESS scores are the scores students obtained in the winter of Grade 1). As we note later in the limitations, this is an important practical barrier to using ACCESS data, as schools would not know students' current-year scores until the end of that year. Although all predictive validity studies are necessarily retrospective, the fact that schools would likely not be able to use current-year ACCESS scores for data-based decision-making is a significant practical limitation.

academic risk (i.e., greater than chance). We did not have an a priori reason to set this criterion at a different value, although there might be reasons to do so if researchers or practitioners believe a different probability criterion constitutes a more accurate indicator of risk.

For models with Grade 1 data, we used Winter M-COMP or R-CBM and/or ACCESS scores to predict Spring MAP scores. For Grades 2–5, we used Fall M-COMP or R-CBM and/or ACCESS scores to predict Spring MAP scores. We were limited to Winter M-COMP/R-CBM data for Grade 1 because schools did not administer Fall M-COMP or R-CBM (instead, they administered early numeracy and early literacy measures, which fall outside the scope of the current study). In Grades 2–5, we used Fall data (rather than Winter, like Grade 1) because schools' universal screening data from the Fall of each year may constitute the first data point of the new year, providing potentially vital information on students' strengths and needs at the beginning of the year. This Fall data point (and, in Grade 1, Winter) helps schools route students into appropriate services to target core mathematics and reading skills. To this end, the first data point for M-COMP and R-CBM during the school year is arguably the most important to examine in terms of predictive validity because this may be the first chance that schools have to conduct normative comparison and route students to subsequent screening gates (for more on multiple-gated screening, see Walker et al., 2014) or intervention services. Thus, it is crucial to understand how this first decision-making point predicts to the end of the year.

Missing Data

Missingness in the current analyses came from a variety of sources. Certain schools did not collect any data on the measures of interest or collected data on only a selection of measures (e.g., some schools did not administer MAP and/or the CBMs of interest in any years or only in one specific cohort year). Additionally, some schools did not report ACCESS scale scores even

though they provided WIDA ELP levels. ACCESS scale scores were utilized given that these scores are more granular measure of ELP than the proficiency scores. Aside from data missing from entire schools or schools within cohorts (10%–25%), remaining missingness (10%–25%) appeared attributable to attrition or other factors (e.g., absence).

Considering the number and types of models within the intended analysis plan (e.g., displaying receiver operating characteristic [ROC] curves), the main body of the paper presents models using complete case analyses. However, complete case analysis typically makes untenable assumptions about missingness (unless the data are missing completely at random, which is unlikely; Enders, 2010). Multiple imputation is one method that can help recover missing at-random data (data that are missing as a result of other observable, measured characteristics; Enders, 2010) that would otherwise be deleted in complete case analysis. We also conducted analyses using multiply-imputed data as a robustness check for the complete case analyses. For imputation, we used predictive mean matching (PMM; Rubin, 1987), which is a semi-parametric method of imputation wherein imputed values are generated based on "donors" with similar values of the predicted missing values (van Buuren, 2018). PMM has a number of advantages, including robustness to different response distributions and imputing only plausible values, and it generally performs well compared to other popular imputation methods (Kleinke, 2017; Vink et al., 2014). The results of analyses with multiple imputation are presented in the Supplemental Materials. Conclusions from the multiple imputation results do not differ substantially from the complete-case analyses presented here.

Data Inspection

Prior to analyses, the distributional properties of each measure to assess the tenability of modeling assumptions for OLS and logistic regression were inspected. Across all years, both the

ACCESS and MAP scores (reading and mathematics) showed relatively univariate normal distributions. However, the CBM data (R-CBM and M-COMP) showed different distributions dependent on the grade level of administration. In the first administration of these measures (Winter of Grade 1), there was a strong right skew to the data indicating a slight floor effect. Between Grades 2–5, the distribution became more normal. Because the CBM measures are used as independent variables, their distribution is of less concern than dependent variables and the distribution of residual errors (homoskedasticity). Nevertheless, highly non-normal predictor data can still compromise the assumptions of generalized linear models (GLMs). We conducted diagnostic checks of the assumptions of our models to determine what model adjustments, if any, needed to be made.

Results

Descriptive Statistics

Descriptive statistics of the predictor and outcome variables are presented in Table 1.

Regression Analyses

For each of the two CBM measures, three block-wise regression models were conducted for each grade level (i.e., 30 models in total). First, each CBM was entered into a regression model on its own. Then a model with only ACCESS predicting MAP scores was conducted. Finally, a model with both the CBM and ACCESS measures predicting MAP outcomes was conducted.

Upon running the initial OLS and logistic models, we performed regression diagnostics to evaluate whether these models met the assumptions of linearity and homoskedasticity and to also assess the presence of high-leverage outliers. Standard OLS regression analyses revealed a number of high-leverage data points that potentially affected both linearity and homoskedasticity

assumptions, necessitating a method to mitigate the data points' leverage. The need to maintain equal sample sizes across nested model comparisons using fit indices (e.g., R^2) precluded the removal of high-leverage data points in some models but not others. Additionally, these outlying data points may have represented qualitatively meaningful data that could be relevant for assessing the adequacy of these screening measures in practice. Thus, deleting these outliers based on leverage alone did not seem a practical solution. Accordingly, we used robust regression using the *robustbase* package in R (Maechler et al., 2020) to estimate linear and logistic models. Robust regression reweights data points to mitigate the influence of high-leverage outliers in estimating residual deviance (e.g., R^2 in linear models), point estimates, and standard errors (Maronna et al., 2019). As a result, high-leverage outliers have reduced weight relative to data that fall closer to the estimated regression line. Model comparison tests (e.g., comparing a model with ACCESS and M-COMP to a model with just M-COMP) to assess model improvement can be conducted similarly to typical regression models; we use a Wald test for this procedure (Maechler et al., 2020).

The R^2 and adjusted R^2 values for each of the M-COMP and R-CBM models across grades are shown in Table 2. We also reported area under the curve (AUC) values for robust logistic regression models as well as sensitivity and specificity values. We only report R^2, AUC, and sensitivity and specificity values as we were more concerned with the unique and shared variance or predictive accuracy of each measure rather than their individual regression coefficients. We discuss the results specific to each domain (reading/mathematics) below. The results of these models using multiply-imputed data are presented in the Supplemental Materials.

Mathematics

The linear regression results show that, in every instance, adding the ACCESS scores to a simple linear model with M-COMP predicting spring MAP mathematics performance resulted in a significant reduction in residual variance. Both the R^2 values and Wald χ^2 tests indicated the significance of this additional explained variance. We observed a similar pattern when adding M-COMP to a simple linear model of ACCESS predicting spring MAP mathematics performance, such that M-COMP explained significant unique variance beyond that of the ACCESS. Although the amount of variance explained differed by grade and sample size, the patterns of unique variance explained among M-COMP and ACCESS is consistent across Grades 1–5.

Reading

The pattern of incremental value of ACCESS observed within mathematics was similar for reading, on the whole. Adding ACCESS to the models with ORF predicting MAP reading resulted in a significant increase in R^2 in all years. In other words, ACCESS uniquely explained a meaningful amount of variance for which ORF does not account.

EL ROC Curves

Although linear regression helps gauge the amount of variance in interim assessment outcomes, it does not necessarily help model the validity of measures in distinguishing at-risk from not at-risk status. The logistic regression and ROC curve analysis serve exactly this function. The procedure for conducting the logistic regression analyses was identical to that of the linear regression analyses (i.e., 30 logistic regression models were run: three for each grade across five grades for both reading and mathematics domains). Again, for brevity, this study does not present the log-odds coefficients for each term in each model. Given that this study's research questions concerned the global predictive capacity of each model rather than the unique

effects of each predictor in each model, it only presents predicted versus observed risk status and the ROC curve analyses.

The AUC, sensitivity, and specificity values for each model are reported in Table 2. The majority of sensitivity/specificity values were largely above chance (i.e., .50). Some models had values closer to .50, suggesting that these models were mixed in their ability to distinguish those at risk from those not at risk when "at-risk" status is operationalized as a ≥ .50 probability of scoring > 1 *SD* below the national MAP norms. Across all grades, we observed an increase in AUC values when including ACCESS on top of M-COMP. Although the Wald tests do not directly correspond to the AUC values themselves, the model comparison tests revealed that adding ACCESS to an M-COMP-only model resulted in a significant decrease in residual deviance in all grades. The reverse, adding M-COMP on top of ACCESS, resulted in less residual deviance decrease in all grades except Grade 4. For ORF, ACCESS and ORF exhibited similar AUC values across all grades (.80–.95). Adding ACCESS in addition to ORF resulted in an improved model across all grades. Adding ORF to an ACCESS model significantly reduced residual deviance in Grades 1–3, but this was not the case in Grade 4 or Grade 5. In addition, Figure 1 displays the ROC curves for the mathematics models in each grade, and Figure 2 displays the ROC curves for the reading models in each grade. These figures visually display the increases in the predicted accuracy when including ACCESS that we observe among the AUC values.

Comparisons with Non-EL Data

To address Research Question 2, robust linear regression analyses were conducted among the non-EL students in the sample for MAP reading and mathematics using ORF and M-COMP,

respectively. The analyses also included ROC analyses using the *roc* function of *pROC*.[3] Thus, there were 20 models in total: one robust logistic and one robust linear regression model for each grade for M-COMP and ORF. Table 3 displays the R^2 and AUC values for each linear or logit model predicting spring MAP achievement or risk status. AUC values were adequate across all years among non-ELs, although they tended to decrease across grades for M-COMP. The ROC curves presented in Figure 3 for each measure in each grade visually display a similar pattern. Descriptive cross-tabulations from both measures showed that there were proportionally very few students scoring below the 1 *SD* cutoff as compared to these proportions among ELs. This leads to less opportunity for students to be misclassified when operationalizing "risk" as a $\geq .50$ probability of scoring >1 *SD* below the national MAP mean. This contributes to the strong AUC values from the logistic regression and very high specificity (and very low sensitivity) values.

To test the predictive validity of the ROC curves for ORF and M-COMP among students classified as ELs and non-ELs, we used the *roc.test* function in the *pROC* package with 10,000 bootstraps to test whether the difference between the ROC curves was 0 (a two-sided test). Across 10 comparisons (Grades 1–5 for MCOMP and ORF), only the Grade 3 MCOMP comparison significantly differed at the $p < .05$ level. All other AUC differences across ELs and non-ELs were not statistically significant at the $p < .05$ level. These results suggest that M-COMP at Grade 3 may possess less predictive capability on its own among ELs as compared to M-COMP among non-ELs. Considering that the ACCESS had significant incremental value in improving AUCs in Grade 3, this suggested that including ELP in M-COMP screening may help recover potentially meaningful differences in the predictive validity of M-COMP between ELs

[3] The *roc* function and a simple logistic regression analysis produce identical ROC curves. Moreover, one robust logistic regression did not converge properly, so we used regular (nonrobust) logistic regression in that case. However, robust and nonrobust models resulted in identical AUCs. Consequently, using the *roc* function produced results identical to all other models.

and non-ELS. Although we identified predicted risk status as having a > .50 predicted probability of scoring below the spring benchmark, this is not the value that optimizes sensitivity and specificity. Among students classified as ELs, however, the probability that optimized sensitivity and specificity fell between .23 and .63 in all CBM-only logit models, suggesting that .50 might be a viable cut-off point for discriminating risk status among ELs when using only CBMs. Among non-ELs, the predicted probability that maximized sensitivity and specificity values fell between .02 and .10. In other words, classification accuracy is most balanced between sensitivity and specificity at probabilities of spring risk status of less than .10. These results suggest that the same CBM measures optimally classify risk status in very different ways given the disproportionate number of ELs scoring >1 *SD* below national MAP means compared to the proportion scoring at that level among non-ELs. We present more information on these probability thresholds in Table S2 of the supplementary materials.

False Discovery Rate Adjustment

Since we conducted a large number (70) of statistical tests related to our hypotheses at the .05 level, we adjusted the false discovery rate using the Benjamini-Hochberg (BH) correction (Benjamini & Hochberg, 1995) using $\alpha = .05$ consistent with What Works Clearinghouse (2020). This adjustment changed the significance of only one hypothesis test, which was the logistic model comparison when adding R-CBM in addition to ACCESS when predicted MAP Reading scores (original $p = .046$). All other hypothesis tests that were already significant at the .05 level remained significant after applying the BH correction.

Discussion

The current study simultaneously considered English proficiency and the predictive validity of R-CBM and M-COMP among ELs. We used multiple analytic techniques to model

relationships between CBMs, ELP, and criterion interim assessment outcomes, including robust linear and logistic regression, as well as ROC curves. The primary finding was that the ACCESS incrementally explained more variance than M-COMP or R-CBM alone, and ACCESS generally improved accuracy in predicting risk status as compared to using only M-COMP or R-CBM. These findings indicated that the ACCESS consistently possessed incremental predictive value above both ORF and mathematics computation in elementary school.

For reading, the results suggested that ORF and ELP explained unique variance in English reading performance at the end of the year, which may be expected given that ORF and ELP both require English language and reading, although in different capacities. These findings are consistent with prior work showing a positive relationship between ORF and reading comprehension among ELs (Quirk & Beem, 2012). Moreover, prior work has established relationships between various forms of English language proficiency and English reading comprehension (Galloway & Uccelli, 2019; Gottardo & Mueller, 2009; Lesaux et al., 2010). However, these findings are not necessarily consistent. For example, whereas Gottardo and Mueller found that both English oral language and word reading predicted comprehension, Lesaux et al. (2010) only found that oral language predicted comprehension.

Importantly, multiple studies have established text reading fluency (i.e., ORF) as a mediator between word reading and oral language and reading comprehension. Kim (2012) provided evidence for this mediating model among Spanish-speaking ELs, and Kim and Wagner (2015) provided further evidence for this among an English-speaking sample. Kim's (2012) results are the most relevant to the current study as the author found that oral language and oral reading fluency uniquely related to reading comprehension, but that oral language did not predict oral reading fluency when controlling for word reading automaticity. Kim (2012) reasoned these

relationships, which countered prior studies (Kim et al., 2011), could be due to ELs not yet possessing the necessary automaticity with English word reading to employ comprehension strategies during ORF tasks. Kim (2012) also noted the discrepant results could be attributable to ELs possessing less English oral language proficiency, and similar to word reading automaticity, a minimum capacity of English oral language might be necessary for it to play a unique role in text reading fluency. Nevertheless, English oral language uniquely predicted reading comprehension, consistent with prior work (Gottardo & Mueller, 2009; Lesaux et al., 2010).

Our findings suggest that ORF (which taps a variety of reading skills related to word reading, prosody, and text comprehension; Kim, 2012; Kim & Wagner, 2015) on its own explains one-third to one-half of the variance in spring MAP reading and that ACCESS explains a non-trivial amount of variance beyond ORF. ELP, as assessed on ACCESS, comprises a variety of comprehension and literacy skills as well as oral and written language. However, as prior work has noted (e.g., Kim, 2012; Kim & Wagner, 2015), ORF taps a unique subset of skills (e.g., word-level reading, text comprehension, automaticity) that links ORF to comprehension, although Kim (2012) showed that the language and word-level predictors of ORF might differ among ELs as compared to English-proficient students. Altogether, our current findings, when combined with prior literature, indicate that ORF captures important dimensions that are useful for predicting reading comprehension, but there are many possible nonoverlapping elements between ORF and ACCESS that require additional attention in research and practice.

Within mathematics, a general measure such as MAP Mathematics likely requires a variety of language and comprehension skills that could potentially be viewed as arising from similar pathways as reading comprehension, which may account for ELP explaining nontrivial variance and predictive accuracy over and above students' computation skills. We posit that this

is the likely reason that the mathematics results generally parallel reading analyses. The math findings also align with previous research indicating that language skills predict math achievement among all students (Chow & Eckholm, 2019; Vukovic & Leasux, 2013a), including language-minoritized groups (Vukovic & Lesaux, 2013b). The limited skill range captured on M-COMP (Christ et al., 2008) further corroborates the incremental value of ACCESS in predicting MAP math. As Christ et al. (2008) noted, the skill range assessed on M-COMP—despite being used as a general outcome measure (GOM)—is unlikely to capture all the necessary elements required for attaining grade-level math proficiency. Procedural and conceptual knowledge are both necessary elements of math cognition (Rittle-Johnson et al., 2015). Although computation provides the necessary algorithmic basis for carrying out mathematics problem solving, conceptual mathematical reasoning might require stronger roots in language than procedures (Vukovic & Lesaux, 2013b). For ELs, greater ELP would likely facilitate access to the mathematics-specific vocabulary and text comprehension skills (in English) necessary to perform higher on an assessment like MAP.

Overall, these results generally supported the hypothesis for Research Question 1 and suggested that screening in both reading and math may benefit from greater consideration of students' ELP levels. However, it is also important to note that the incremental increase in predictive accuracy (based on AUC values) after adding ACCESS to models was smaller at lower grade levels, particularly for ORF. One possibility for this pattern is that ORFs' capacity to predict reading comprehension changes based on the underlying skills required at different developmental levels of reading (Kim & Wagner, 2015). A similar process might be the case for M-COMP, as the linguistic complexity of math procedures and concepts increases over time (and would be reflected on MAP), whereas M-COMP continues to focus on grade-level calculation

only. However, given the different sample sizes across grades and the differing number of students falling below our "risk" criterion at each grade, it is difficult to determine to what this variation in incremental value of ACCESS should be attributed. Nevertheless, variation across grades in ACCESS's incremental value should be a consideration when interpreting our results.

A second finding was that R-CBM and M-COMP exhibited adequate predictive capacities across language groups. However, it is difficult to directly compare these samples of students given the large sample size differences and different prevalence rates of being at risk. Therefore, this conclusion should be considered with significant caution and should be corroborated with larger samples of ELs. Nevertheless, the small EL samples here are likely to mirror the small EL demographic in many school districts. As a result, conclusions are based on a practically relevant sampling scenario that many applied researchers and practitioners are likely to encounter and need to find ways to handle in a robust manner. These sampling barriers necessitate the use of rigorous data-based decision-making practices to ensure equity in intervention provision and identification (Albers et al., 2013; Albers & Martinez, 2015).

Implications for Research

These results provide some insight into the use of CBMs to screen for academic performance; however, additional research is needed to fully understand the academic skills being measured. For example, R-CBM assesses reading fluency as "words correct per minute." However, other constructs, such as comprehension and prosody, are also necessary for adequate reading (Kim et al., 2021) and are not fully measured in R-CBM. This may have implications for the validity of these measures as more complex skills are taught and assessed in later grades (again, it is possible the lower predictive validity of ORF in Grades 4 and 5 may be a product of

this). With the inclusion of ACCESS, the predictive validity improved, perhaps given the more comprehensive focus on multiple reading and language skills across both measures.

Additionally, because language has a unique relationship with reading and math, the interaction and measurement of language, reading, and math skills necessitates further attention, especially for ELs. Previous research has shown a significant association between reading and math skills (Chen & Chalhoub-Deville, 2016; Grimm, 2008; Thurber et al., 2002) as well as between language and mathematics skills (Chow & Ekholm, 2019; Peng et al., 2020; Vukovic & Lesaux, 2013a, 2013b). Moreover, complex language negatively impacts comprehension of math application problems (Abedi & Lord, 2001). The high language demands of potentially complex math application problems threaten the validity of math skills assessments for students from language-minoritized backgrounds (Abedi & Gandara, 2007; Martiniello, 2008; Sireci et al., 2008; Wolf & Leon, 2009). A greater understanding of the acquisition of these academic skills and their complex relationships with one another can be used to inform the refinement of CBMs, the frequency and timing of administration of these measures for students classified as ELs, and instructional supports for these students.

A secondary line of research that arises from these findings involves the predictive validity of CBM and language proficiency measures with different EL subgroups. EL populations within schools are heterogenous, despite the fact that research tends to focus on Spanish-speaking students, and native Spanish-speaking ELs generally represent the majority of ELs in the United States (Hussar et al., 2020). Therefore, additional research is needed to understand the use of these measures with students classified as ELs among whom there is substantial variation in native languages.

Finally, it may be beneficial to investigate the validity of CBMs across ELP levels (e.g., Kim et al., 2016) and with varying levels of ELP skills. For example, the ACCESS measures multiple English language domains. Although composite variables tend to explain the majority of variance across subdomains (Canivez, 2013), understanding how different levels of subcomponents of language may moderate relationships between screeners and outcomes help with tailoring multiple-gated screening decisions (Walker et al., 2014) and intervention provision. For example, among a variety of language skills, language syntax skills most robustly predicted math performance in elementary school (Chow & Ekholm, 2019), and English oral language was more strongly related to conceptually-focused mathematics tasks than procedural tasks (Vukovic & Lesaux, 2013b). Previous research found that the predictive performance of ORF was equal across three ELP categories (Kim et al., 2016). However, additional research is needed to support these previous findings and understand how they can be informative for practice.

Implications for Practice

This study's primary finding suggests the importance of reviewing ELP data in addition to academic universal screening data to determine additional assessment or intervention services for EL students. This implication aligns with Albers and Martinez's (2015) proposed RTI model, which encourages school personnel to consider ELP levels and language acquisition factors at each stage of RTI, including universal screening. Given school psychologists' training in data-based decision-making, school psychologists may be best suited to initiate conversations about ELP considerations. To integrate academic and language data in the decision-making process for EL students, Albers et al. (2013) recommended comparing EL's academic screening data with

other EL's data using local norms collected from regular administrations of academic universal screeners at that school or district.

Although CBMs might be predictively valid for ELs, schools must ensure the utility of academic screening data by considering the appropriateness of the screener's constructs and its outcome data in relation to classroom instruction (Glover & Albers, 2007). School psychologists may consult with curriculum leaders to determine how to best assess relevant skill development and may use single skill-specific probes to monitor students' response to instruction. These practices that can target more specific academic strengths and needs of ELs can be greatly supplemented by the existing literature on the relationships between text reading fluency, math computation, and general mathematics and reading achievement. Given that in both the mathematics and reading literature there is considerable shared variance with many aspects of language, school psychologists should be considering ELP data as part of existing screening practices.

In terms of using screening data for decision-making (e.g., deciding to route students into an additional screening gate or provide intervention), our findings, when combined with other extant research, suggests that no single data point should be considered alone and that the constructs underlying each measure need to be carefully considered. In particular, these measurement aspects of screening should be contextualized within instruction in the specific school context. Students' grade level, ELP level, and M-COMP scores should be interpreted in terms of what are reasonable grade-level expectations based on curricular standards, normative comparisons, and local comparisons. This line of questioning should follow from all aspects of data collection among ELs (and non-ELs) in order to give the most meaning to screening data, given that the development of native and English language proficiencies interacts in a variety of

ways with assessment and instructional content (Albers & Martinez, 2015). Combining screening and ELP data can provide greater context as to how instruction and formative assessment can promote ELs' attainment of grade-level standards (Richardson et al., 2020).

On a broader level, this research can have implications for policies that define the use of assessment and instruction within an RTI model. These findings highlight the need to advocate for additional attention, in the form of increased funding and research, on academic screening in general, the use of screeners associated with math and reading skills specifically, and populations of students from language-minoritized backgrounds, including students classified as ELs. **The sensitivity and specificity of M-COMP and R-CBM among ELs in the current study is consistent with prior work (Keller-Margulis et al., 2008; Kilgus et al., 2014; Shapiro et al., 2015). Under the current definition of academic risk, these screeners tend to show similar predictive patterns as what has been observed in prior studies focusing on the general population of students. It is important to note, however, that ELs' lower performance on MAP means that the base rate among ELs is inherently higher, which is one contributor to why the current study finds diagnostic accuracy consistent with prior work (prior to accounting for ACCESS scores). If the cut score were set higher (e.g., the 40th percentile, such as Balu et al., 2015), the specificity of the measure would be very poor (the opposite of non-ELs, for whom sensitivity is very poor due to a low base rate). This reemphasizes the need to establish norms among ELs (Albers & Martinez, 2015) as well as within the local context, as cut scores have been shown to moderate diagnostic accuracy of R-CBM (Kilgus et al., 2014). Moreover, understanding the local context and local needs of ELs may help reduce erroneous decision making, such as the base rate fallacy (Bowes et al., 2020).**

Screening is a critical mechanism of prevention that ideally helps identify the strengths and needs of students at the school population-level, thereby facilitating targeted services to promote core grade-level skills (Clemens et al., 2016). To promote the equity of universal supports, screening practices should be equally effective for all eligible students. Increased attention is needed in policy and advocacy of these universal preventative supports that can increase the efficiency and effectiveness of service provision across tiers of support.

Limitations

Given this study's findings, it is important to describe several limitations. As aforementioned, the small EL samples compared to non-ELs leaves the analyses possibly insufficient to conclusively make comparisons across language groups given that these comparisons are so heavily weighted towards non-ELs. Future studies would benefit from larger samples of ELs, although our analyses that utilized multiple imputation may have allowed us to possibly recover some power lost to complete-case analysis, and the multiple imputation results follow a similar pattern to the complete case results. We combined data across cohorts of students, which means that the same students may appear in multiple grades. Since we conducted analyses separately by grade, there are not any repeated observations of students within the models, although there may be some dependence of observations across models.

Additionally, the cutoff for determining "at risk" students was limited in scope. Relying on a > 1 *SD* cut-score is likely insufficient to determine risk status and should be corroborated with additional qualitative and quantitative information. Moreover, the predictive accuracy analyses could be sensitive to small changes to this criterion, especially among the smaller EL sample sizes when changes in the proportions of at-risk students would carry more weight. The characteristics of the data, such as outliers and missing data from individual students as well as

entire schools, also impact the study's conclusions. Although supplemental analyses suggest that the core findings using robust regression hold even in light of multiply imputing missing data, the results should be interpreted cautiously given that outliers and the amount of missing data may bias results in the multiple imputation analyses. Additionally, these data were collected for a larger project that ended approximately six years ago, and classroom instruction, assessment practices, and RTI practices likely have changed in that time. Perhaps most pertinent to practice is the temporality of measurements used in this study. Except for Grade 1 (which uses winter CBM), this study utilized fall CBM measures to predict spring outcomes; however, ACCESS assessments are typically administered in late fall or early winter. The analyses can be seen as retrospective in nature, given that schools tend not to have fall CBM and up-to-date ACCESS data available simultaneously. Prior year ACCESS performance tends to inform EL status in the subsequent year, so data-based decision-making using the models presented may not accurately reflect the available data when critical service provision decisions are made.

Conclusion

Within an RTI model, academic universal screeners (e.g., CBMs) are commonly used to inform important educational decisions, including the need for interventions and supports for students considered at risk for academic difficulty. To ensure that equitable educational decisions are made, academic screeners must provide accurate data regarding student abilities and skill acquisition. Therefore, it is imperative to ensure that these tools are valid for use with all students, including students classified as ELs. For ELs in particular, ELP levels may be used in addition to CBM data to enhance accuracy in predicting student risk status and explaining variance in spring assessment scores.

References

Abedi, J., & Gándara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement Issues and Practice, 25*(4), 36–46. doi:1.1111/j.1745-3992.2006.00077.x

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*, 219–234. doi:1.1207/s15324818ame1403_2

Albers, C. A., Kenyon, D. M., & Boals, T. J. (2009). Measures for determining English language proficiency and the resulting implications for instructional provision and intervention. *Assessment for Effective Intervention, 34*, 74–85. doi:1.1177/1534508408314175

Albers, C. A., & Martinez, R. (2015). *Promoting success with English language learners: Best practices for RTI.* Guilford.

Albers, C. A., Mission, P. L., & Bice-Urbach, B. (2013). Considering diverse learner characteristics in problem-solving assessment. In R. Brown-Chidsey & K. J. Andren (Eds.), *Assessment for intervention: A problem-solving approach* (2nd ed., pp. 101–122). Guilford.

Allsopp, D. H., Farmer, J. L., & Hoppey, D. (2016). Preservice teacher education and response to intervention within multi-tiered systems of support: What can we learn from research and practice? In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.), *Handbook of response to intervention: The science and practice of multi-tiered systems of support* (2nd ed., pp. 143–158). Springer.

Baker, S. K., & Good, R. (1995). Curriculum-based measurement of English reading with bilingual Hispanic students: A validation study with second-grade students. *School Psychology Review, 24*, 561–578.

Ballard & Tighe Publishers. (1991*). IDEA Language Proficiency Tests (IPT) – English*. Author.

Balu, R., Zhu, P., Doolittle, F., Schiller, E., Jenkins, J, & Gersten, R. (2015). *Evaluation of response to intervention practices for elementary school reading*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U. S. Department of Education. https://files.eric.ed.gov/fulltext/ED560820.pdf

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57*(1), 289–300.

Bowes, S. M., Ammirati, R. J., Costello, T. H., Basterfield, C., & Lilienfeld, S. O. (2020). Cognitive biases, heuristics, and logical fallacies in clinical practice: A brief field guide for practicing clinicians and supervisors. *Professional Psychology: Research and Practice*, 51(5), 435–445. http://dx.doi.org/10.1037/pro0000309

Burns, M. K., Appleton, J. J., & Stehouwer, J. D. (2005). Meta-analytic review of responsiveness-to-intervention research: Examining field-based and research-implemented models. *Journal of Psychoeducational Assessment, 23*, 381–394. doi:1.1177/073428290502300406

Canivez, G.L. (2013). Incremental criterion validity of WAIS–IV factor index scores: Relationships with WIAT–II and WIAT–III subtest and composite scores. *Psychological Assessment*, *25*, 484–495. doi: 1.1037/h0088996

Center for Applied Linguistics. (2014). *Annual technical report for ACCESS for ELLs English language proficiency test, Series 301, 2012-2013 administration* (WIDA Consortium Annual Technical Report No. 9). Author.

Center for Applied Linguistics. (2017). *Annual technical report for ACCESS for ELLs 2.0 online English language proficiency test, Series 400, 2015-2016 administration* (WIDA Consortium Annual Technical Report No. 12A).

Chen, F., & Chalhoub-Deville, M. (2016). Differential and long-term language impact on math. *Language Testing*, *33*, 577–605. doi:1.1177/0265532215594641

Christ, T. J., Scullin, S., Tolbize, A., & Jiban, C. L. (2008). Implications of recent research: Curriculum-based measurement of math computation. *Assessment for Effective Intervention, 33*, 198–205. doi:1.1177/1534508407313480

Chow, J. C., & Ekholm, E. (2019). Language domains differentially predict mathematics performance in young children. *Early Childhood Research Quarterly*, *46*, 179–186. doi: 1.1016/j.ecresq.2018.02.011

Clemens, N. H., Keller-Margulis, M. A., Scholten, T. S., & Yoon, M. (2016). Screening assessment within a multi-tiered system of support: Current practices, advances, and next steps. In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.), *Handbook of response to intervention: The science and practice of multi-tiered systems of support* (2nd ed., pp. 187–214). Springer.

CRESST. (n.d.). *English Language Proficiency Assessment for the 21st Century (ELPA21)*. CREST. https://cresst.org/elpa21/

Data Recognition Corporation. (n.d.). *Language Assessment Scales (LAS)*. Author. https://laslinks.com/

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition*, *44*, 1–42.

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education*, *37*, 184–192. doi:/1.1177/00224669030370030801

Enders, C. A. (2010*). Applied missing data analysis*. Guilford.

Every Student Succeeds Act of 2015 (ESSA), Pub. L. No. 114-95 § 114 Stat. 1177 (2015-2016).

Fien, H., Nelson, N. J., Smolkowski, K. Kosty, D., Pilger, M., Baker, S. K., & Smith, J. L. M. (2021). A conceptual replication study of the enhanced core reading instruction MTSS-reading model. *Exceptional Children*, *87*(3), 265–288. doi: 1.1177.0014402920953763

Fletcher, J., & Vaughn, S. (2009). Response to intervention: Preventing and remediating academic difficulties. *Child Development Perspectives, 3*, 30–37. doi:1.1111/j.1750-8606.2008.00072.x

Fuchs, L. S., Wang, A. Y., Preacher, K. J., Malone, A. S., Fuchs, D., & Pachmayr, R. (2021). Addressing challenging mathematics standards with at-risk learners: A randomized controlled trial on the effects of fractions intervention at third grade. *Exceptional Children*, *87*(2), 163–182. doi: 1.1177/0014402920924846

Galloway, E P., & Uccelli, P. (2019). Examining developmental relations between core academic language skills and reading comprehension for English learners and their peers. *Journal of Educational Psychology*, *111*(1), 15–31. http://dx.doi.org/1.1037/edu0000276

Geary, D. C. (2010). Missouri longitudinal study of mathematical development and disability. *Understanding Number Development and Difficulties*: *BJEP Monograph Series*, *II*(7), 31–49. doi: 1.1348/97818543370009X12583699332410

Gersten, R., Jordan, N. C., & Flojo, J. R. (2005). Early identification and interventions for students with mathematics difficulties. *Journal of Learning Disabilities*, *38*(4), 293–304. doi: 1.1177/00222194050380040301

Glover, T. A. & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45,* 117–135. doi:1.1016/j.jsp.2006.05.005

Gottardo A., & Mueller, J. (2009). Are first- and second-language factors related in predicting second-language reading comprehension? A study of Spanish-speaking children acquiring English as a second language from first to second grade. *Journal of Educational Psychology*, *101*(2), 330–344. https://doi.org/ 1.1037/a0014320

Gottlieb, M., Cranley, M.E., & Cammilleri, A. (2007). *Understanding the WIDA English language proficiency standards.* Board of Regents of the University of Wisconsin System.

Grimm, K. J. (2008). Longitudinal associations between reading and mathematics achievement. *Developmental Neuropsychology, 33*, 410–426. doi:1.1080/87565640801982486

Harry, B., Klingner, J., Delpit, L., & Artiles, A. (2014). *Why are so many minority students in special education? Understanding race and disability in schools*. Teachers College Press.

Hosp, J. L., Huddle, S., Ford, J. W., & Hensley, K. (2016). Learning disabilities/special education. In S. R. Jimerson, M, K. Burns, & A. M. VanDerHeyden (Eds.), *Handbook of response to intervention: The science and practice of multi-tiered systems of support* (2nd ed., pp. 43–58). Springer.

Hussar, B., Zhang, J., Wang, X., Wang, K., Hein, S., Diliberti, M,. . .Purcell, S. (2020). *The condition of education 2020* (NCES 2020-144). U.S. Department of Education. National Center for Education Statistics. https://nces.ed.gov/pubs2020/2020144.pdf

Individuals with Disabilities Education Improvement Act, 20 U.S.C. § 140. (2004).

January, S. A., Van Norman, E. R., Christ, T. J., Ardoin, S. P., Eckert, T. L., & White, M. J. (2018). Progress monitoring in reading: Comparison of weekly, bimonthly, and monthly assessments for students at risk for reading difficulties in grades 2-4. *School Psychology Review*, *47*(1), 83–94. doi: 1.17105/SPR-2017-0009.V47-1

Lesaux, N. K., Crosson, A. C., Kieffer, M. J., & Pierce, M. (2010). Uneven profiles: Language minority learners' word reading, vocabulary, and reading comprehension skills. *Journal of Applied Developmental Psychology*, *31*(6), 475–483. https://doi.org/1.1016/j.appdev.201.09.004

Keller-Margulis, M. A., Shapiro, E. S., & Hintz, J. M. (2008). Long-term diagnostic accuracy of curriculum-based measures in reading and mathematics. *School Psychology Review*, *37*(3), 374–39. doi: 1.1080/02796015.2008.12087884

Kieffer, M. J., & Thompson, K. D. (2018). Hidden progress of multilingual students on NAEP. *Educational Researcher, 47*, 391–398. doi:1.3102/0013189X1877777740

Kilgus, S. P., Methe, S. A., Maggin, D. M., & Tomasula, J. L. (2014). Curriculum-based measurement of oral reading (R-CBM): A diagnostic test accuracy meta-analysis of evidence supporting use in universal screening. *Journal of School Psychology*, *52*(4), 377–405. https://doi.org/10.1016/j.jsp.2014.06.002

Kim, J. S., Vanderwood, M. L., & Lee, C. Y. (2016). Predictive validity of curriculum-based measures for English language learners at varying English proficiency levels. *Educational Assessment, 21*, 1–18. doi: 1.1080/10627197.2015.1127750

Kim, Y. -S. (2012). The relations among L1 (Spanish) literacy skills, L2 (English) language, L2 text reading fluency, and L2 reading comprehension for Spanish-Speaking ELL first grade students. *Learning and Individual Differences*, *22*, 690–700.

Kim, Y.-S. G., Quinn, J. M., & Petscher, Y. (2021). Reading prosody unpacked: A longitudinal investigation of its dimensionality and relation with word reading and listening comprehension for children in primary grades. *Journal of Educational Psychology*, *113*(3), 423–445. https://doi.org/1.1037/edu0000480

Kim, Y.-S., & Wagner, R. K. (2015). Text (oral) reading fluency as a construct in reading development: An investigation of its mediating role for children from grades 1 to 4. *Scientific Studies of Reading*, *19*(3), 224–242. https://doi.org/1.1080/10888438.201

Kim, Y. -S., Wagner, R. K., & Foster, L. (2011). Relations among oral reading fluency, silent reading fluency, and reading comprehension: A latent variable study of first-grade readers. *Scientific Studies of Reading*, *15*, 338–362.

Kleinke, K. (2017). Multiple imputation under violated distributional assumptions: A systematic evaluation of the assumed robustness of predictive mean matching. *Journal of Educational and Behavioral Statistics*, *42*(4), 371–404. https://doi.org/1.3102/1076998616687084

LeFevre, J., Fast, L., Skwarchuk, S.-L., Smith-Chant, B. L., Bisanz, J., Kamawar, D., & Penner-Wilger, M. (2010). Pathways to mathematics: longitudinal predictors of performance. *Child Development*, *81*(6), 1753–1767. doi: 1.1111/j.1467-8624.201.01508.x

Lesaux, N. K., Crosson, A. C., Kieffer, M. J., & Pierce, M. (2010). Uneven profiles: Language minority learners' word reading, vocabulary, and reading comprehension skills. *Journal of Applied Developmental Psychology*, *31*(6), 475–483. doi: 1.1016/j.appdev.201.09.004

Maculaitis, J, D. (n.d*.). Maculaitis Assessment of Competencies Test of English Language Proficiency*. Questar. http://www.questarai.com/assessments/english-language-proficiency-assessments/

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T…di Palma, M.A. (2020). *robustbase: Basic robust statistics R package version 0.93-6*. http://CRAN.R-project.org/package=robustbase.

Maronna, R. A., Martin, R. D., Yohai, V. J., & Salibián-Barrera, M. (2019). *Robust statistics: Theory and methods (with R)* (2nd ed.). John Wiley & Sons, Inc.

Martiniello, M. (2008). Language and the performance of English language learners in math word problems. *Harvard Educational Review, 78*, 333–368. doi:1.17763/haer.78.2.70783570r1111t32

Méndez, L. I., Hammer, C. S., Lopez, L. M., & Blair, C. (2019). Examining language and early numeracy skills in young Latino dual language learners. *Early Childhood Research Quarterly*, *46*, 252–261. doi: 1.1016/j.ecresq.2018.02.004

MetriTech, Inc. (n.d.). *Language proficiency test series.* Author.

Morgan, P. L., Farkas, G., Cook, M., Strassfeld, N.M., Hillemeier, M. M., Pun, W. K…Schussler, D. L. (2018). Are Hispanic, Asian, Native American, or language-minority children overrepresented in special education? *Exceptional Children*, *84*(3), 261–279. doi: 1.1177/0014402917748303

Muyskens, P., Betts, J., Lau, M. Y., & Marston, D. (2009). Predictive validity of curriculum-based measures in the reading assessment of students who are English language learners. *The California School Psychologist*, *14*, 11–21. doi: 1.1007/BF03340947

Nelson, G., & Powell, S. R. (2018). A systematic review of longitudinal studies of mathematics difficulty. *Journal of Learning Disabilities*, *51*(6), 523–539. doi: 1.1177.0022219417714773

National Association of School Psychologists. (2020). *The professional standards of the National Association of School Psychologists.* Author.

Northwest Evaluation Association. (2004). *Reliability and validity estimates: NWEA achievement level tests and measures of academic progress.* Author.

Northwest Evaluation Association. (2011). *Technical manual: For measures of academic progress (MAP) and measures of academic progress for primary grades (MPG).* Author.

Northwest Evaluation Association. (2015). *2015 Measures of academic progress normative data*. Author.

O'Bryon, E., & Rogers, M. (2010). Bilingual school psychologists' assessment practices with English language learners. *Psychology in the Schools, 47*, 1018–1034. doi:1.1002/pits.20521

Pearson Education, Inc. (2012). *AIMSweb technical manual.* Author.

Peng, P., Lin, X., Ünal, Z. E., Lee, K., Namkung, J., Chow, J. & Sales, A. (2020). Examining the mutual relations between language and mathematics: A meta-analysis. *Psychological Bulletin*, *146*(7), 595–634. doi: 1.1037/bul0000231

Poncy, B. C., Skinner, C. H., & Axtell, P. K. (2005). An investigation of the reliability and standard error of measurement of words read correctly per minute using curriculum-based measurement. *Journal of Psychoeducational Assessment, 23,* 326–338. doi:1.1177/073428290502300403

Quirk, M., & Beem, S. (2012). Examining the relations between reading fluency and reading comprehension for English language learners. *Psychology in the Schools*, *49*(6), 539-553. https://doi.org/10.1002/pits.21616

R Core Team. (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.

Richardson, R. D., Rocconi, L. M., & Crewdson, M. A. (2020). Evaluating English learner progress in reading: How much growth can we expect? *School Psychology Review*, *49*(4), 480-492. https://doi.org/10.1080/2372966X.2020.1787080

Roberts, G., & Bryant, D. (2011). Early mathematics achievement trajectories: English-language learner and native English-speaker estimates, using the Early Childhood Longitudinal Survey. *Developmental Psychology, 47*, 916–930. doi: 1.1037/a0023865

Roberts, G., Mohammed, S. S., & Vaughn, S. (2010). Reading achievement across three language groups: Growth estimates for overall reading and reading subskills obtained with the Early Childhood Longitudinal Survey. *Journal of Educational Psychology, 102*(3), 668–686. https://doi.org/1.1037/a0018983

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek F., Sanchez, J-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77–85. doi:1.1186/1471-2105-12-77

Robinson-Cimpian, J. P., Thompson, K. D., & Umansky, I. M. (2016). Research and policy considerations for English learner equity. *Policy Insights from the Behavioral and Brain Sciences*, *3*(1), 129–137. https://doi.org/10.1177/2372732215623553

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, Inc.

Samson, J. F., & Lesaux, N. K. (2009). Language-minority learners in special education: Rates and predictors of identification for services. *Journal of Learning Disabilities, 42*, 148–162. doi:1.1177/0022219408326221

Shapiro, E. S., Dennis, M. S., & Fu, Q. (2015). Comparing computer adaptive and curriculum-based measures of math in progress monitoring. *School Psychology Quarterly*, *30*(4), 470–487. https://doi.org/10.1037/spq0000116

Sireci, S. G., Han, K. T., & Wells, C. S. (2008). Methods for evaluating the validity of test scores for English language learners. *Educational Assessment, 13*, 108–131. doi:1.1080/10627190802394255

Sullivan, A. L. (2011). Disproportionality in special education identification and placement of English language learners. *Exceptional Children*, *77*(3) 317–334.

Sullivan, A. L., & Bal, A. (2013). Disproportionality in special education: Effects of individual and school variables on disability risk. *Exceptional Children*, *79*(4), 475–494.

Thurber, R. S., Shinn, M. R., & Smolkowski, K. (2002). What is measured in mathematics tests? Construct validity of curriculum-based mathematics measures. *School Psychology Review, 31*, 498–513.

Umansky, I. M., Thompson, K. D., & Diaz, G. (2017). Using an ever-English learner framework to examine disproportionality in special education. *Exceptional Children*, *84*(1), 76–96. doi: 1.1177/0014402917707470

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). CRC Press.

Van Norman, E. R., & Nelson, P. M. (2021). The importance of growth in oral reading fluency to predict performance on high-stakes assessments among students receiving supplemental instruction. *Journal of Applied School Psychology*, *37*(1), 1–15. doi: 1.1080/15377903.202.1772432

Vanderheyden, A., & Burns, M. (2005). Using curriculum-based assessment and curriculum-based measurement to guide elementary mathematics instruction: Effect on individual and group accountability scores. *Assessment for Effective Intervention, 30*, 15–31. doi:1.1177/073724770503000302

Vanderwood, M. L, Tung, C. T., & Checca, C. J. (2014). Predictive validity and accuracy of oral reading fluency for English learners. *Journal of Psychoeducational Assessment*, *32*(3), 249–258. doi: 1.1177/0734282913502937

Vink, G., Frank, L. E., Pannekoek, J., & van Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, *68*(1), 61–90. https://doi.org/1.1111/stan.12023

Vukovic, R. K., & Lesaux, N. K. (2013a). The relationships between linguistic skills and arithmetic knowledge. *Learning and Individual Differences*, *23*, 87–91. doi: 1.1016/j.lindif.2012.1.007

Vukovic, R. K. & Lesaux, N. K. (2013b). The language of mathematics: Investigating the ways language counts for children's mathematic development. *Journal of Experimental Child Psychology, 115,* 227–224. doi: 1.1016/j.jecp.2013.02.002

What Works Clearinghouse. (2020). *What Works Clearinghouse procedures handbook, version 4.1*. Institute of Education Sciences, National Center for Educational Evaluation and Regional Assistance, U.S. Department of Education. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Procedures-Handbook-v4-1-508.pdf

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.

WIDA. (2020). *WIDA Consortium member states and territories.* https://wida.wisc.edu/memberships/consortium

Wiley, H. I., & Deno, S. L. (2005). Oral reading and maze measures as predictors of success for English learners on a state standards assessment. *Remedial and Special Education, 26*, 207–214. doi:1.1177/07419325050260040301

Wolf, M. K., & Faulkner-Bond, M. (2016). Validating English language proficiency assessment uses for English Learners: Academic language proficiency and content assessment

performance. *Educational Measurement: Issues and Practice, 35*(2), 6–18. doi:1.1111/emip.12105

Wolf, M. K., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment, 14*, 139–159. doi:1.1080/10627190903425883

Table 1

Measure Descriptive Statistics

Grade	Measure	*N*	*M*	*SD*	% At Risk
1	MCOMP	100	17.20	12.86	
	ORF	143	31.52	27.38	
	MAP Math	142	169.86	15.62	40%
	MAP Reading	143	169.48	14.47	34%
	ACCESS	143	293.44	20.94	
2	MCOMP	125	10.42	7.95	
	ORF	125	38.49	27.10	
	MAP Math	125	182.38	12.91	34%
	MAP Reading	124	179.96	13.27	31%
	ACCESS	125	316.86	19.77	
3	MCOMP	111	12.23	9.61	
	ORF	111	55.75	39.48	
	MAP Math	111	191.77	12.60	38%
	MAP Reading	111	188.02	15.27	33%
	ACCESS	111	343.03	12.26	
4	MCOMP	71	11.11	7.98	
	ORF	71	63.56	37.19	
	MAP Math	71	198.07	15.09	55%
	MAP Reading	71	191.59	12.77	37%
	ACCESS	71	348.55	22.77	
5	MCOMP	77	6.06	4.82	
	ORF	76	77.08	37.21	
	MAP Math	77	201.44	13.70	57%
	MAP Reading	77	194.22	14.20	51%
	ACCESS	77	355.40	21.35	

Note. At risk defined as scoring 1 *SD* below the national normative mean in 2015. M-COMP = Mathematics Computation; ORF = Oral Reading Fluency; MAP Math = Measures of Academic Progress – Math; MAP Reading = Measures of Academic Progress – Reading; ACCESS = ACCESS for ELLs English language proficiency measure.

Table 2

Results of Regression Models Predicting MAP Achievement (Linear) or Risk Status (Logistic)

Grade (M-COMP *n*/**ORF *n***)			Outcome: MAP Mathematics			Outcome: MAP Reading		
			M-COMP	ACCESS	Both	ORF	ACCESS	Both
Grade 1 (100/**143**)	Linear	R^2	.582	.552	.716*^	.434	.666	.685*^
		Adj. R^2	.578	.547	.710	.430	.664	.681
	Logistic	AUC	.867	.896	.918*^	.851	.886	.901*^
		Sens/Spec.	(.33/.96)	(.77/.87)	(.85/.87)	(.58/.81)	(.73/.88)	(.77/.87)
Grade 2 (125/**125**)	Linear	R^2	.253	.547	.607*^	.546	.624	.695*^
		Adj. R^2	.247	.543	.600	.543	.621	.690
	Logistic	AUC	.768	.856	.878*^	.887	.893	.913*^
		Sens/Spec.	(.60/.83)	(.65/.89)	(.67/.89)	(.71/.87)	(.61/.87)	(.68/.87)
Grade 3 (111/**110**)	Linear	R^2	.304	.491	.564*^	.501	.741	.768^*
		Adj. R^2	.298	.486	.556	.497	.738	.764
	Logistic	AUC	.765	.804	.836*^	.934	.922	.949*^
		Sens/Spec.	(.57/.84)	(.53/.93)	(.55/.88)	(.78/.89)	(.70/.93)	(.78/.93)
Grade 4 (71/**71**)	Linear	R^2	.256	.427	.485*^	.510	.592	.653^*
		Adj. R^2	.245	.419	.469	.502	.586	.642
	Logistic	AUC	.706	.794	.806*	.870	.859	.893*
		Sens/Spec.	(.74/.50)	(.67/.75)	(.69/.72)	(.69/.84)	(.54/.93)	(.65/.89)
Grade 5 (77/**76**)	Linear	R^2	.224	.534	.600*^	.372	.659	.667*
		Adj. R^2	.213	.528	.589	.363	.655	.657
	Logistic	AUC	.735	.821	.853*^	.801	.910	.921*
		Sens/Spec.	(.82/.36)	(.84/.64)	(.82/.76)	(.71/.71)	(.84/.84)	(.92/.82)

Note. For both linear and logistic models, robust Wald χ^2 nested model comparison tests assessed the decrease in robust residual deviances for each additional term compared to the model with only a single term (CBM or ACCESS). M-COMP = Mathematics Computation; ORF = Oral Reading Fluency; ACCESS = ACCESS for ELLs English language proficiency measure.

*$p < .05$ compared to only CBM, ^$p < .05$ compared to only ACCESS, *^ $p < .05$ compared to both CBM and ACCESS.

Table 3

R^2, AUC, and Sensitivity/Specificity Values for non-EL M-COMP and ORF Regressions

Grade (MCOMP n /**ORF** *n*)	MAP Mathematics (Predictor: M-COMP)			MAP Reading (Predictor: ORF)		
	R^2	AUC	Sensitivity/ Specificity	R^2	AUC	Sensitivity/ Specificity
Grade 1 (1202/**1199**)	.458	.922	.24/.99	.461	.905	.11/1^
Grade 2 (1539/**1534**)	.350	.825	.07/1*	.458	.922	.24/.99
Grade 3 (1465/**1465**)	.378	.877	.13/1*	.429	.918	.33/1*
Grade 4 (1424/**1423**)	.313	.806	.13/1^	.374	.872	.13/1*
Grade 5 (1415/**1409**)	.356	.807	.13/1^	.391	.892	.17/.99

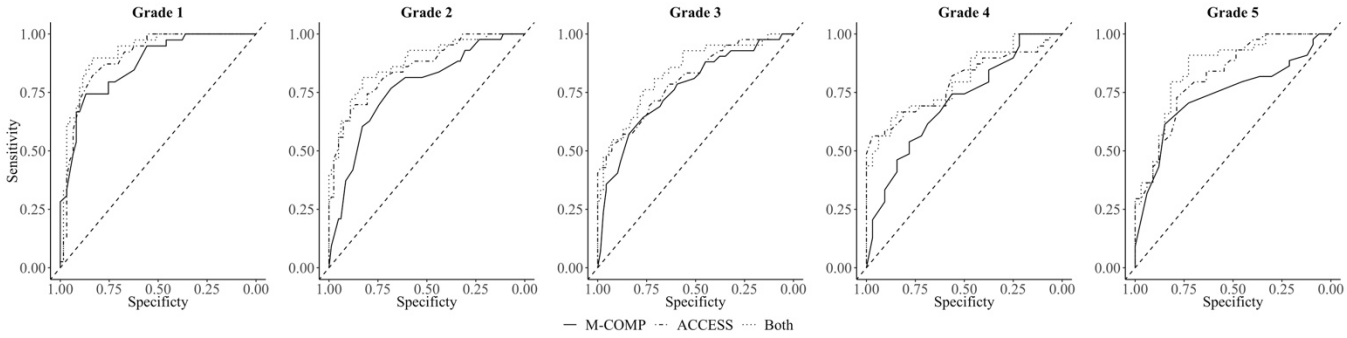
Note. M-COMP = Mathematics Computation; ORF = Oral Reading Fluency; AUC = Area Under the Curve. ORF sample sizes are bolded. R^2 values are from models with MAP achievement as the outcome; AUC and sensitivity/specificity values are from models with binary MAP risk status as the outcome. *Specificity rounded to 1, ^Specificity equaled 1.

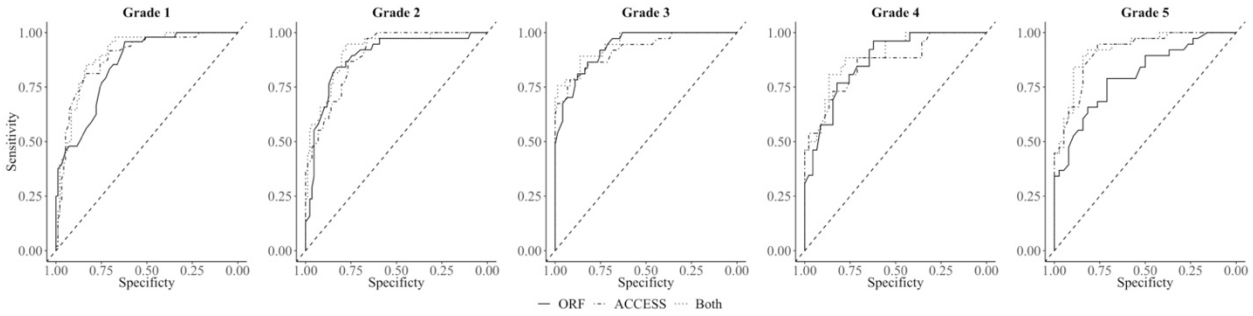
Figure Captions

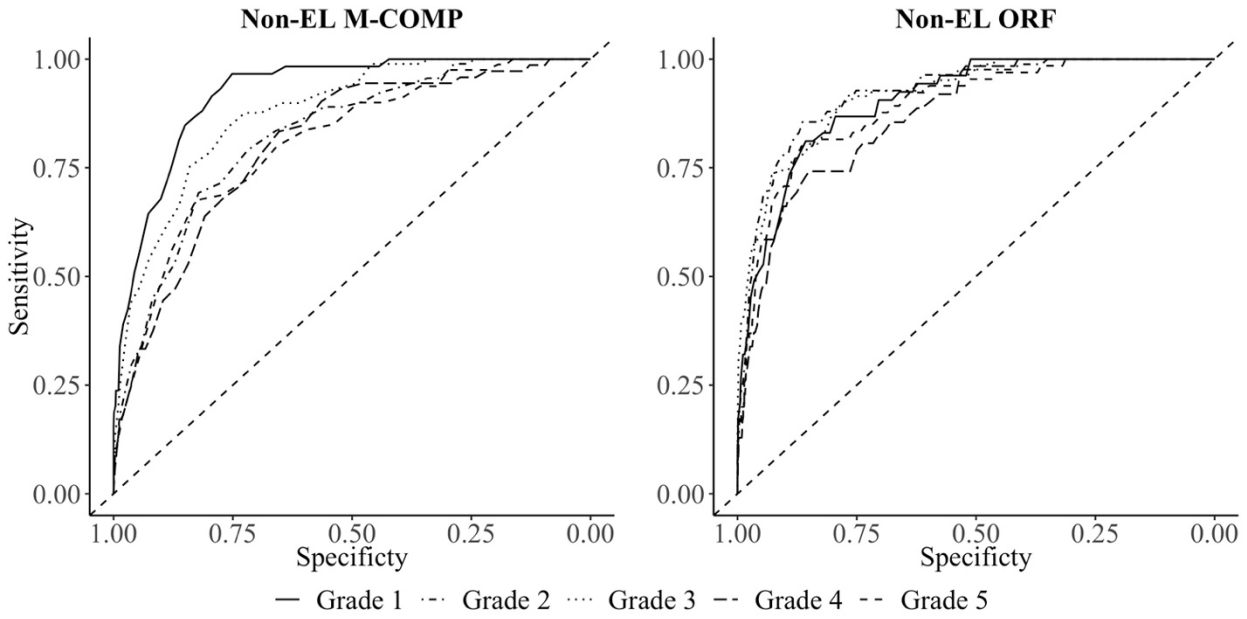
Figure 1. ROC curves for each M-COMP and/or ACCESS logistic regression model predicting spring mathematics at-risk status by grade.

Figure 2. ROC curves for each ORF and/or ACCESS logistic regression model predicting spring reading at-risk status by grade.

Figure 3. M-COMP and ORF ROC curves among non-ELs by grade.


Grade 1
Grade 2
Grade 3
Grade 4
Grade 5
Sensitivity
Specificty
1.00
0.75
0.50
0.25
0.00
M-COMP
ACCESS
Both


Grade 1
Grade 2
Grade 3
Grade 4
Grade 5
Sensitivity
1.00
0.75
0.50
0.25
0.00
Specificty
ORF
ACCESS
Both


Non-EL M-COMP
Non-EL ORF
Sensitivity
Specificty
1.00
0.75
0.50
0.25
0.00
Grade 1
Grade 2
Grade 3
Grade 4
Grade 5

Supplemental Materials for "Predicting Interim Assessment Outcomes Among Elementary-Aged ELs using M-COMP, R-CBM, and English Proficiency"

Within this Supplemental Materials document, we provide several additional analyses to supplement our main results that were presented in the main paper. Due to missingness on our English language proficiency (ELP) variable as well as missing curriculum-based measures and outcome assessment data (i.e., Measures of Academic Progress), we conducted a number of supplemental analyses to test the sensitivity of our primary results to multiple imputation of missingness. We also discuss the optimal cut points of CBM-only diagnostic accuracy models.

Analyses with Multiple Imputation

As reported in the main text, we conducted follow-up analyses using multiple imputation with predictive mean matching (PMM; Rubin, 1987). We used the *mice* package (van Buuren & Groothuis-Oudshoorn, 2011) in R (R Core Team, 2021) to impute 100 datasets with 50 iterations each within each grade level among ELs. Among students who had an ACCESS Overall Composite scale score (as mentioned in the main text, not all students with an ACCESS ELP level had the corresponding Overall Composite scale score), we imputed missingness on the independent and dependent variables within each model. Within our imputation procedure, we included ACCESS scale scores as well as ELP levels, fall (where applicable), winter, and spring M-COMP and ORF; both fall and spring MAP mathematics and reading; and demographics (students' school, binary special education status, and race/ethnicity [including gender prohibited imputation convergence]). The target variables that were imputed were (a) fall (or winter for Grade 1) M-COMP and ORF and (b) spring MAP (reading and mathematics). Following imputation, we created the same MAP risk indicator using the criteria described in the main text. Table S1 below displays R^2 and area under the curve (AUC) estimates that are pooled across

imputed datasets using the *pool.r.squared* function in *mice* to obtain R^2 and adjusted R^2 estimates. AUC values were obtained by manually computing the mean of AUC values from each logistic regression models estimated on each of the 100 datasets.

We also provide Akaike Information Criterion (AIC) values for each of the models, which helps further assess the relative performance of the models that include both CBM and ACCESS to the models with only CBM or only ACCESS. AIC values have no absolute metric so they must be interpreted in relative terms, and models with different variables cannot be compared (i.e., we cannot compare CBM-only models to ACCESS-only models). Thus, the purpose of the AIC is to provide another measure of incremental value of ACCESS in addition to CBM in the linear and logistic models. Burnham et al. (2011) noted that changes in differences in AIC values of greater than approximately 14 give evidence in favor of the model with the smaller AIC, whereas difference values below 14 suggest increasing inconclusiveness regarding the favorable model as difference values approach zero. Based on the changes in the AIC values from CBM-only to CBM plus ACCESS models, we generally find strong evidence for the favorability of including ACCESS on top of M-COMP or R-CBM among both logistic and linear regression models.

Generally, the pattern of results was similar to the complete case analyses in the main paper for the diagnostic accuracy analyses (i.e., logistic regressions); however, there was more variability in the R^2 estimates (presumably because of the amount of missing data that were imputed). It is important to note that a significant portion of the missing data was due to some schools systematically not administering certain M-COMP, ORF, or MAP measures. This creates more uncertainty in the imputation of the missing data, and thus the results presented with multiple imputation should be interpreted with significant caution.

Table S1

Model Fit Indices from Linear and Logistic Regression Models (Estimates Pooled from Multiple Imputation)

			Mathematics			Reading		
			M-COMP	ACCESS	Both	R-CBM	ACCESS	Both
Grade 1	Linear	R^2	.527	.609	.721	.381	.700	.717
		Adj. R^2	.525	.607	.718	.378	.698	.714
		AIC	1696.267	1655.381	1583.018	1730.841	1572.515	1561.618
	Logistic	AUC	.874	.893	.930	.843	.912	.919
		AIC	198.198	185.016	153.280	218.372	171.446	164.953
Grade 2	Linear	R^2	.296	.628	.669	.532	.667	.737
		Adj. R^2	.292	.626	.666	.530	.666	.734
		AIC	1691.647	1554.697	1531.681	1627.827	1554.551	1506.090
	Logistic	AUC	.774	.879	.898	.894	.891	.928
		AIC	244.782	182.405	173.854	176.211	170.593	145.060
Grade 3	Linear	R^2	.297	.542	.604	.419	.696	.715
		Adj. R^2	.293	.540	.601	.416	.694	.712
		AIC	1473.503	1392.973	1367.227	1485.348	1363.519	1353.454
	Logistic	AUC	.766	.790	.829	.893	.910	.936
		AIC	220.011	202.791	190.310	153.910	135.268	119.941
Grade 4	Linear	R^2	.119	.470	.489	.469	.668	.699
		Adj. R^2	.111	.466	.481	.465	.666	.694
		AIC	1155.83	1087.543	1084.682	1082.890	1018.798	1007.656
	Logistic	AUC	.663	.754	.771	.872	.880	.911
		AIC	172.181	156.606	153.523	125.195	116.721	106.311
Grade 5	Linear	R^2	.172	.592	.630	.464	.687	.709
		Adj. R^2	.166	.589	.625	.460	.685	.704
		AIC	1149.536	1050.171	1038.408	1100.437	1024.816	1016.413
	Logistic	AUC	.693	.783	.815	.812	.888	.896
		AIC	183.378	154.021	148.956	150.331	119.667	116.687

Note. AIC = Akaike's Information Criterion, AUC = area under the curve.

Optimal Cut Points for CBM-Only Logistic Regression Models

In Table S2 below we present the predicted probabilities from simple logistic regression models that maximize both sensitivity and specificity values (i.e., the predicted probabilities that maximize distinguishing those actually at risk [sensitivity] from those actually not at risk [specificity]). Table 2 indicates that the optimal predicted probabilities are significantly different between ELs and non-ELs; optimal probabilities for ELs are noticeably higher. Scores associated with a very low probability of predicted risk correspond to more accurate discrimination of at-risk and not at-risk students among non-ELs, whereas higher CBM scores (those corresponding to probabilities between .23–.63) more accurately discriminate risk status among ELs. In other words, ELs must evidence higher CBM scores than non-ELs for their risk status to be accurately classified.

Table S2

Sensitivity and Specificity of Optimal Thresholds in Probability Units

CBM	Grade	EL Status	Threshold	Specificity	Sensitivity
M-COMP	1	EL	.524	.869	.744
		Non-EL	.024	.752	.966
	2	EL	.336	.683	.767
		Non-EL	.095	.822	.692
	3	EL	.502	.841	.571
		Non-EL	.053	.761	.843
	4	EL	.618	.781	.538
		Non-EL	.041	.653	.833
	5	EL	.625	.848	.614
		Non-EL	.092	.828	.675
ORF	1	EL	.233	.621	.958
		Non-EL	.093	.857	.811
	2	EL	.397	.826	.842
		Non-EL	.064	.864	.855
	3	EL	.444	.877	.811
		Non-EL	.025	.788	.890
	4	EL	.414	.822	.769
		Non-EL	.069	.850	.742
	5	EL	.072	.866	.800
		Non-EL	.424	.711	.789

References

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodal inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*(1), 23–35. doi: 1.1007/s00265-010-1029-6

R Core Team. (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* Wiley.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://www.jstatsoft.org/v45/i03/