Heterogeneity in Early Mathematics Screening: Investigating the Role of Intervention Effects on Screening Accuracy

by

Christopher Ives

A dissertation accepted and approved in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in School Psychology

Dissertation Committee:

Gina Biancarosa, Chair

Laura Lee McIntyre, Advisor

Ben Clarke, Core Member

Emily Tanner-Smith, Institutional Representative

University of Oregon

Summer 2023

© 2023 Christopher Ives

DISSERTATION ABSTRACT

Christopher Ives

Doctor of Philosophy in School Psychology

Title: Heterogeneity in Early Mathematics Screening: Investigating the Role of Intervention Effects on Screening Accuracy

This study explores the heterogeneity in screening accuracy of the Assessing Student Proficiency in Early Number Sense (ASPENS) across schools within the context of a randomized control trial (RCT) for Fusion, a first-grade early math intervention. Students were assigned to one of three conditions: a business-as-usual (BAU) control group, a two-student Fusion group (2-Fusion), and a five-student Fusion group (5-Fusion). Two research questions were addressed: 1) To what extent does the observed screening accuracy of ASPENS meaningfully differ between students randomly assigned to the Fusion intervention conditions compared to the BAU condition?; and 2) To what extent is heterogeneity in screening accuracy reduced when is ASPENS is administered concurrently with its criterion, rather than at different times of the year? Data were analyzed using generalized linear mixed models to jointly model sensitivity and specificity at the participant level, using the 20th percentile on the Test of Early Mathematics Ability – 3rd Edition as the reference criterion.

As hypothesized, findings indicated that specificity was significantly affected by treatment conditions, with the 2-Fusion condition exhibiting lower specificity than the BAU condition. 5-Fusion also demonstrated lower specificity than BAU, but this difference was not statistically significant. Furthermore, heterogeneity in screening accuracy across treatment groups was no longer evident when assessments were administered concurrently. The findings of this study underscore the challenges of prognostic screening frameworks and have implications

for the use of publisher-recommended cut-scores, the development and validation of academic screening measures, and guiding best practices in utilizing screening assessments within multi-tiered systems of support.

CURRICULUM VITAE

NAME OF AUTHOR: Christopher Ives

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED

University of Oregon, Eugene

DEGREES AWARDED:

Doctor of Philosophy, School Psychology, 2023, University of Oregon
Bachelor of Music, Violin Performance, 2017, University of Oregon

AREAS OF SPECIAL INTEREST:

Educational Measurement
Universal Screening in Multi-tiered Systems of Support
Early Academic Intervention Supports

PROFESSIONAL EXPERIENCE:

School Psychology Intern, Springfield Public Schools, 2022 – 2023

Graduate Teaching Assistant, Academic Programming, College of Education, University of Oregon, Summer 2021

Graduate Teaching Assistant, Education Studies, College of Education, Spring 2021

Graduate Research Assistant, Center on Teaching and Learning, University of Oregon, 2019 – 2022

Research Assistant, Center for Improvement of Child and Family Services, Portland State University, Summer 2017

Research Assistant, Center for Improvement of Child and Family Services, Portland State University, Summer 2017

Research Assistant, Oregon Center for Optics, University of Oregon, Summer 2017

Research Assistant, Prevention Science Institute, University of Oregon, 2016 – 2017

Educational Assistant, B.E.S.T. Afterschool Program, Eugene School District 4J, 2017

Special Education Assistant, Spring Creek Elementary, Eugene School District 4J, 2016 – 2017

Special Education Assistant, Peninsula Union School District, Humboldt County Office of Education, Summer 2016

GRANTS, AWARDS, AND HONORS

Dynamic Measurement Group Award, University of Oregon, 2020

PUBLICATIONS:

Furjanic, D., Ives, C., Fainstein, D., Kennedy, P., Biancarosa, G. (In press). Investigating Changes in Oral Reading Fluency Growth During the COVID-19 Pandemic. *Elementary School Journal.*

ACKNOWLEDGEMENTS

I wish express my gratitude to Dr. Gina Biancarosa, for her assistance in the preparation of this manuscript, years of nurturing mentorship, and her critical role in fostering my passion for statistics and educational measurement. I would also like to thank Laura-Lee McIntyre, Ben Clarke, and Emily-Tanner Smith for their thoughtful feedback and encouragement during the dissertation process. I am grateful to Patrick Kennedy for his guidance and open-door policy as I developed my statistical programming skills at the Center for Teaching and Learning, and for first introducing me to a receiver operating characteristic analysis. In addition to my family's unwavering support, I thank my father, for instilling in me the confidence to tackle difficult challenges in statistics and programming, and for my mother, whose career as an educator exposed me to a meaningful pursuit to devote my passion and energy to. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through grants R324A090341 and R324A160046 to the Center on Teaching and Learning at the University of Oregon.

TABLE OF CONTENTS

LIST OF TABLES

I: INTRODUCTION

Despite several decades of research and nationwide efforts dedicated to improving students' literacy and math skills, a majority of students in the United States continue to perform below national proficiency benchmarks on math and reading assessments (NAEP, 2022). In response to these protracted and unresolved concerns, schools have increasingly adopted multi-tiered systems of support (MTSS), which integrate evidence-based instructional practices within a prevention-oriented framework (Balu, 2015; Gersten et al., 2009; Samuels, 2011). MTSS utilized a tiered service delivery model to strategically allocate supplementary resources to students based on their identified level of need. The theoretical foundations for MTSS have their origins in decades of research highlighting the need for early prevention and intervention to address academic problems before they become more challenging to remediate (Juel, 1988; Kame'enui & Carnine, 1998; McCardle et al., 2001; Scarborough, 1998). However, successful implementation of this model is contingent on effective early identification practices, as underpinned using accurate measurement tools.

Universal Screening in MTSS

Prevention-focused service delivery models like MTSS rely on screening assessments to identify students at risk for future academic difficulties. The results of these screening assessments are then used to facilitate the provision of appropriate interventions and supplemental support for at-risk students (Petscher et al., 2011; Fuchs et al., 2004). As a result, screening assessments must be highly accurate in distinguishing students who are not on track towards proficiency, both to support efficient allocation of school resources and to provide timely opportunities to intervene before students’ academic difficulties become intractable.

Typically, universal screening is conducted three times per year with the entirety of a school’s student population. Such a frequent assessment schedule is enabled by screening assessments’ brevity and ease of administration, which make their use feasible across a school building (Glover & Albers, 2007). However, despite being brief and convenient, they must be highly accurate in distinguishing students that are not on track towards proficiency before their academic difficulties become actualized.

Misidentification of non-cases (i.e., false positives) can lead to unnecessary intervention services, resulting in inefficient resource allocation and undermining the primary rationale for implementing tiered supports (Jenkins et al., 2007). Conversely, misidentification of cases (i.e., false negatives) can prevent schools from providing timely intervention services to students in need, ultimately hindering efforts to improve at-risk students' academic trajectories by the end of the school year.

Validity of CBMs for Screening Purposes

Curriculum-based measures (CBMs) are the most common form of screening assessments and are used for several additional purposes, including progress monitoring, program evaluation, and survey-level assessment (Deno, 1985; Fuchs et al., 2004; Kilgus et al., 2014). Because of their multiple applications, CBMs necessitate a more complex validity argument. Under Kane’s (2013) argument-based validity framework, the interpretations or intended use of test scores require individual evaluations rather than wholly ascribing validity to an assessment. Conventional criterion-related validity can support the basis for interpreting CBM scores relative to a particular construct or trait value (e.g., math computation, reading fluency), but further diagnostic accuracy evidence is necessary to justify their use as a screener (Kilgus et al., 2014).

Screening decisions, as implemented in most settings, conclude with a dichotomous prediction of high vs. low risk. Thus, validity evidence in support of an assessment's use for screening must indicate that it reliably differentiates students by their risk for academic difficulties within a dichotomous interpretive framework. Furthermore, diagnostic accuracy evaluations, as with any assessment evaluation or research hypothesis, reflect the testing of specific hypotheses or interpretations in particular subpopulations and settings (Gambino, 2018). Hence, it is a misnomer that screening accuracy does not require the same investigation into generalizability as other aspects of test design – a misnomer currently expressed in the National Center for Intensive Intervention's specific omission of review processes related to sample representativeness or bias analyses for classification accuracy in its tool chart rubric for academic screening tools (National Center on Intensive Intervention [NCII], 2020).

Several meta-analyses have examined the correlational evidence, or criterion-related validity, of CBMs relative to a criterion measure (January & Klingbeil, 2020; Reschly et al., 2009; Yeo, 2010), but these are not evidence for a screening tool's ability to reliably distinguish academic risk or need for intervention supports in a MTSS framework. As previously mentioned, an adequate validity investigation into using CBMs for screening purposes must examine its discriminative properties within a dichotomous identification framework (e.g.,-typical achievement vs. at-risk students). As of yet, only one meta-analysis has investigated variations in CBM screening accuracy (Kilgus et al., 2014). The dearth of meta-analyses or explicit investigations into variability in academic screening accuracy is concerning, given that large differences between studies are relatively common in diagnostic accuracy research which often cannot be attributed to mere chance (Macaskill et al., 2010).

Diagnostic Accuracy

Although screening assessments and the procedures for establishing their decision thresholds can be constructed in various ways, the methods for evaluating screening accuracy are relatively consistent. Statistically, these methods are parallel to the evaluation of diagnostic or classification systems more broadly, in which a dichotomous status outcome is generated (Smolkowski & Cummings, 2015). Thus, "screening accuracy," "diagnostic accuracy," and "classification accuracy" are all methodologically synonymous terms, with screening accuracy only differing in the context in which it is used.

Diagnostic accuracy evaluation, as with any assessment evaluation or research hypothesis, reflects testing specific hypotheses or interpretations in particular populations and settings (Gambino, 2018). Diagnostic accuracy is typically measured by estimating a test's precision in distinguishing those with the condition (i.e., cases) from those without the condition (i.e., non-cases). The process for differentiating cases from non-cases must be appropriate and accurate, given that this differentiation forms the basis on which the accuracy of the screening assessment is judged. Hence, it is traditional for cases and non-cases to be identified using a gold-standard reference assessment.

When screening for academic difficulties, the condition of interest is typically represented by performance below a certain threshold (e.g., 20th percentile) on an EOY norm-referenced test of broader academic achievement. CBM screening tools yield a dichotomous prediction as to whether a student is likely to perform above or below this threshold on the criterion assessment (i.e., at-risk, not at-risk). Conventionally, students that fall below the threshold for academic proficiency are referred to as “truly at-risk,” and students who meet or exceed that threshold are referred to as “truly not at-risk” (Catts et al., 2015; Klingbeil et al., 2015; Vanderheyden, 2013). Such language can quickly become opaque, as “risk” implicates a future occurrence and alludes to statistical probabilities that are conceptually ill-defined. Thus, to the extent possible, I will henceforth refer to *truly* at-risk students as “cases” of academic difficulty and *truly not* at-risk students as “non-cases” (Gambino, 2006).

Misclassification

Two types of error are possible under a dichotomous identification framework: false positives and false negatives. False positives (FP) refer to students identified as cases on the screening assessment but surpassed the threshold for academic difficulty on the criterion assessment. False negatives (FN) refer to students who met the definition of academic difficulty on the criterion assessment but were “missed” or designated as non-cases on the screening assessment. Conversely, accurate identification occurs when there is classification agreement between the screening and criterion assessment. True positives (TPs) and true negatives (TNs) represent instances in which a screener accurately identified cases and non-cases of academic difficulty, respectively. This resulting array of four possible categorizations is summarized in Table 1.

Table 1.

Confusion matrix coding method for screening accuracy

Screening Result	Status on Criterion Assessment	
	Cases	Non-Cases
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

Regardless of how the condition truly manifests, diagnostic accuracy systems assume that those with and without the condition are distinct populations. In academic screening, FNs characterize the screening tools' accuracy in correctly identifying students within the *academic difficulty* population, and FPs indicate accuracy among *typically achieving* students. The proportions of FPs and FNs relative to their respective populations are used to calculate the most common indices of diagnostic accuracy – sensitivity and specificity.

Sensitivity represents the proportion of individuals in the academic difficulty population correctly identified by the screener (i.e., 1 – false-positive rate), while specificity represents the proportion of individuals in the typically-achieving population that were accurately identified (i.e., 1 – false-negative rate). A tool's overall accuracy is measured using its receiver operating characteristic (ROC) curve, which plots a measure's sensitivity versus its false-positive rate (FPR; 1 – specificity) across all possible cut-scores. The area under the curve (AUC), or *c* statistic, provides an overall summary of the screening tool's discriminatory power, with values ranging from 0 to 1.0. Interpretively, the AUC also indicates the probability that a randomly selected case would obtain a lower score on the screening assessment than a randomly selected control (Cook, 2007; Hanley & McNeil, 1982). An AUC value of 0.50 indicates that a tool lacks any discriminatory power and that the probability of distinguishing a case from a control approximates random chance. Conversely, an AUC of 1.0 indicates that a case will always obtain a lower score than a control, thus demonstrating perfect discrimination. When evaluating screening tools, AUC values exceeding .90 suggest excellent accuracy, values between .70 and .90 are useful for some purposes, and values below .70 are indicative of poor accuracy (Swets, 1988).

As previously mentioned, sensitivity and specificity represent a screening accuracy's ability to distinguish cases and non-cases at a particular cut-score. In an applied context, these accuracy indices are some of the most consequential, as they more closely represent expected accuracy in an operational screening framework where cut scores are utilized to support decision-making. Within the context of academic screening, there are sensitivity and specificity threshold minimums for what is considered an acceptable screening tool, though there is some variation in these criteria. For example, the National Center on Intensive Intervention (NCII) recommends a minimum sensitivity and specificity of .80 (National Center on Intensive Intervention [NCII], 2020); however, some advocate for more stringent criteria, with a minimum sensitivity of .90 (Clemens et al., 2001, Compton et al., 2006; Jenkins, 2003; Klingbeil et al., 2021).

It is important to note that, even with equivalent sensitivity and specificity values, misclassification errors are not equally distributed. That is, a screening tool with perfectly balanced sensitivity and specificity will not produce a 1:1 ratio of FPs to FNs because these metrics are proportionally referencing different populations. For example, with a sample base rate of .20 and sensitivity and specificity values of .80, a screening tool would produce approximately four FPs for each FN. These calculations are illustrated in Table 2, assuming a population size of 100.

Depending on the cut-score selected on the screening tool, sensitivity can be increased, but at the cost of decreasing specificity. Appropriately balancing these metrics is an important component of cut score selection and implicates a value judgement as to whether sensitivity or specificity should be privileged. It is commonly affirmed that sensitivity should be privileged over specificity, since there are more dire consequences associated with failing to provide an at-risk student with supports, rather than providing unnecessary support to a student that is on track (Clemens et al., 2011; Jenkins, 2003). However, decisions to favor sensitivity or specificity are arguably more complex when considering screening systems at the school level, especially when considering the prevalence of academic difficulties in a particular context. This is because, as previously mentioned, sensitivity and specificity alone will not indicate the proportion of FP to FNs. Rather, because they represent proportions of statistically distinguished populations (i.e., cases vs. non-cases), the system-level burdens of privileging sensitivity are exaggerated in a low base rate school, since the accompanying sacrifice to specificity results in significantly more misclassifications as a proportion of the large population of typically achieving students.

Table 2.

Illustrative example of sensitivity and specificity calculations

Screening Result	Status on Criterion Assessment		
	Cases	Non-Cases	Row Totals
Positive	TP = 64	FP = 16	TP + FP = 80
Negative	FN = 4	TN = 16	FN + TN = 20
Column Totals	TP + FN = 68	FP + TN = 32	*N* = 100

Note. TP = True positive; FP = False positive; FN = False negative; TN = True negative. Sensitivity = 64/(64 + 16) = 0.8; Specificity = 16/(4 + 16) = 0.8; Base Rate = 20/(80 + 20) = .20.

Issues in Screening Accuracy

Whereas some indices (i.e., positive predictive power, negative predictive power) are discouraged because they are sample-dependent, sensitivity and specificity are widely relied on because they are thought to be population-level statistics and can be treated as properties of the test itself (Johnson et al., 2009; Kleingbeil et al., 2018; Petscher et al., 2011; VanDerHeyden, 2011). In purely their calculations, this is true. Because sensitivity and specificity separately describe a cut scores accuracy in regard to cases and non-cases, respectively, they do not inherently depend on the proportions of cases to non-cases reflected in a particular sample or subpopulation.

However, the commonly held assumption that sensitivity and specificity are sample-independent is contradicted by diagnostic accuracy research that has found they covary with the prevalence of the condition, or base rate, in particular subpopulations (Brenner & Gefeller, 1997; Cook, 2007; Leeflang et al., 2009, 2013). In some scenarios, sensitivity and specificity can express similar levels of variation due to base rate as other indices previously discouraged in the educational literature for this very reason (Brenner & Gefeller, 1997). If certain diagnostic accuracy indices are indeed deemed inappropriate for summarizing a tool's diagnostic accuracy due to their reliance on base rate (Smolkowski & Cummings, 2015; Swets, 1988), it would follow that the field's concern should persist if there are similar vulnerabilities to the accuracy indices presently in common use.

Notwithstanding the issues of diagnostic accuracy indices, several meta-analyses have examined the correlational evidence, or criterion-related validity, of CBMs relative to an criterion measure (January & Klingbeil, 2020; Reschly et al., 2008; Yeo, 2010). However, these are not evidence for a screening tool's ability to reliably distinguish academic risk or need for intervention supports in a MTSS framework. Kilgus et al. (2014) offer the only meta-analysis of CBM screening accuracy, focusing on CBM Oral Reading (R-CBM) in Grades 3-8. Although they found that R-CBM cut scores performed adequately across studies, the specific cut-scores used were inconsistent. For example, Kilgus et al. (2014) report that Grade 3 BOY cut-scores predicting a criterion with six-months lag ranged from 45 to 83 words per minute (WPM; *M* = 61.80, *SD* = 12.56). Though some variation in optimum cut scores is to be expected due to different forms and assessment systems, this wide range of cut scores depicts significant heterogeneity in the level of R-CBM performance constituting risk. Although one of the central goals of their meta-analysis was to explore heterogeneity in screening accuracy, they implicitly analyzed the best performing cut-score within each study sample (i.e., mixed threshold analysis), meaning the optimum cut-scores were selected post-hoc and were allowed to freely vary across studies. As a result, they were unable to describe how the performance of a particular benchmark varies across settings, such as if a school or district were to utilize publisher-recommended cut-scores. Instead, the results of their meta-analysis summarize the performance of R-CBM in identifying students at-risk for reading difficulties when using cut-scores optimized after the criterion assessment has already been conducted. A more naturalistic study of diagnostic accuracy would evaluate a screening tool's ability in predicting performance on the criterion using cut-scores identified a priori.

If generalized sensitivity and specificity estimates meet the minimum criterion for acceptable screening accuracy when permitting significant variation in cut scores, as was found by Kilgus et al. (2014), it raises questions as to how consistent screening accuracy remains when applying a singular nationwide cut-score across settings, as is commonly done in current practice. In other words, if studies were required to select meaningfully different cut scores to achieve an acceptable balance of sensitivity and specificity, it would follow that enforcing a consistent cut score would likely lead to an imbalance across most of the samples.

Among studies that have cross-validated vendor- or publisher-recommended cut scores with a new sample, many found inadequate screening accuracy for the identified cut scores in both math (e.g., Klingbeil et al., 2018, 2021) and reading (e.g., Hintze et al., 2003; Johnson et al., 2009; Klingbeil et al., 2015). In summary, the heterogeneity of screening accuracy when using publisher-recommended cut-scores remains unknown for any CBM across the domains of math or reading, though many authors advocate for the use of local cut-scores due to the inadequacy of publisher-recommended cut-scores for their samples (Keller-Margulis et al., 2008; Klingbeil et al., 2012; Nelson et al., 2017; Patton et al., 2014; Thomas & January, 2019).

Screening accuracy is commonly evaluated with a tool's sensitivity and specificity – statistical indices which summarize an assessment's ability to discriminate between two populations (e.g., typically-achieving students vs. at-risk students) at a specific cut score. Sensitivity and specificity are used to validate screening assessments and identify optimum cut scores in applied settings. Critically, some research has found that these indices can fail to generalize to local contexts or subpopulations (Brenner & Gefeller, 1997; Cook, 2007; Leeflang et al. 2009, 2013). Such variance in screening accuracy is further underscored by the many studies advocating for the use of local cut-scores due to the inadequacy of publisher-recommended cut-scores for their samples (Keller-Margulis et al., 2008; Klingbeil et al., 2012; Nelson et al., 2017; Patton et al., 2014; Thomas & January, 2019). Considering universal cut-scores are still widely used, and sensitivity and specificity indices are used as broad-brush descriptors of a screener's performance, there continues to be a need for investigation into why these variations in screening accuracy are observed.

Potential Vulnerabilities in Academic Screening Accuracy

A critical issue in the generalizability of screening accuracy is the lag time introduced between the screening and criterion assessment, such as when researchers and publishers validate a beginning of year (BOY) screener against and end of year (EOY) criterion. Such a lag introduces the “treatment paradox.” The treatment paradox refers to instances in which patients’ screening results are used to initiate treatment prior to the administration of the diagnostic criterion. As a result, the prediction generated from the screening instrument is effectively disrupted due to intervention. Rutjes et al. (2006), in their meta-analysis of biases within diagnostic accuracy studies, found no significant effects of treatment on diagnostic accuracy; however, their analysis was not regarding school-based academic screening, where the relevance of this phenomenon is arguably more compelling.

Consider that screening validation studies do not withhold interventions from students that are at-risk for EOY academic difficulties. Instead, screening evaluations are typically conducted in a naturalistic setting, meaning educators are often responding to screening data and providing interventions to students in need of them. This has become particularly true as MTSS systems and universal screening procedures have become more widespread, as indicated by majority of states with active screening requirements related to dyslexia (Youman & Mather, 2018). When cut scores for risk are identified in naturalistic contexts, the meaning of risk becomes intwined with the school support systems in the validation sample. That is, when academic supports are exercised between the administration of the screening and reference assessment, the resulting “optimized” cut scores do not indicate the likelihood of a performing below proficiency expectations in the *absence* of intervention. Rather, the predictive cut-scores represent the likelihood of sub-proficient performance *in spite of* the existing academic supports. Such information arguably has more limited actionable value to educators and risks undermining their morale and the face validity of MTSS systems. For example, an important premise of MTSS systems is that, on average, academic interventions should be sufficient to accelerate progress among lower-performing students to achieve proficiency by EOY. If screening tools’ true definition of academic risk refers to the likelihood of reaching proficiency despite intervention, then the intensity of support must exceed that in the validation sample to effectively mitigate risk. Furthermore, instructing educators to respond to screening results with a relatively similar intervention protocol to that in the validation sample, where efforts failed to mitigate risk among other students with the same scores, creates a situation where teachers and paraprofessionals are set up to administer inadequate levels of support.

Regardless of whether students' abilities change due to intervention efforts or for other reasons, when there is a delay in time between the screening assessment and reference assessment, individuals can "migrate" from the typically achieving population to the academic difficulty population, and vice versa. For example, a student that indeed was a member of the math difficulty population could have sufficiently benefited from instruction and intervention supports that they no longer met the criterion for math difficulty when assessed months later on the criterion assessment. Henceforth, this phenomenon is termed *positive risk migration*, which is a distinct phenomenon from measurement error. Positive risk migration refers to a student whose position on the normative, latent continuum of math ability sufficiently increases between the screening and reference assessment, such that they cross the threshold distinguishing typically achieving students from those with academic difficulties. Conversely, *negative risk migration* would refer to students that are truly in the typically achieving population at the time of screening, but whose normative ability has fallen between the screening and reference assessments and are now classified to be a member of the math difficulty population.

The potential causes for positive and negative risk migration are numerous. They may include the quality of core instruction, curricular alignment, provision of supplemental supports, individual fluctuations in development or academic growth, and any other contextual variables that can influence change in the latent condition of interest. With prognostic screening validation, where the screener is used to predict the students' future status, it is not possible to distinguish students that were misidentified due to screening error from those that demonstrated risk migration. Considering differential growth in math skills has been observed based on disability status, English-language proficiency status, socioeconomic background, and other demographic characteristics (Scammacca et al., 2020; Wei et al., 2012), the adoption of a prognostic screening framework can introduce bias that is misinterpreted as screening tool error given that they assume a certain degree of growth that is implicitly presumed to be unbiased in their validation sample.

Present Study

The goal of this study is to examine whether the screening accuracy of an early mathematics screener varies across schools within the context of a randomized control trial (RCT) for an early math intervention. That is, I will investigate whether screening accuracy is moderated by student- and school-level characteristics, including the provision of an evidence-based. Specifically, I will be evaluating the screening accuracy of the Assessing Student Proficiency in Early Number Sense (ASPENS; Clarke et al., 2011) assessment in predicting later performance on the Test of Early Mathematics Ability – 3rd Edition (TEMA-3; Ginsburg & Baroody, 2003) among schools participating in a randomized control trial (RCT) of the Fusion Math Intervention.

Research Questions

This study intends to answer the following research questions:

Research Question 1: To what extent does the observed screening accuracy of ASPENS meaningfully differ between students randomly assigned to the Fusion Math Intervention compared to a business-as-usual (BAU) condition?

Research Question 2: To what extent is heterogeneity in screening accuracy reduced when the ASPENS and TEMA-3 have been administered concurrently, rather than at different times of year (i.e., BOY vs. EOY)?

The present study will borrow from meta-analytic methods of diagnostic accuracy research but will treat schools as the unit of analysis. That is, each school will conceptually represent a study of the diagnostic accuracy of ASPENS within its unique context and subpopulation. This novel approach will provide a practically relevant summary of ASPENS screening accuracy, given that it reflects a meaningful division (i.e., schools) of the educational landscape in which screening assessments are used. Naturally, this investigation has the potential to expose contexts in which screening accuracy is poorer than others. For example, it is expected that variation in screening accuracy will be particularly exacerbated when there is a delay between the screening and criterion assessment, which introduces an opportunity for school-specific environmental variables (e.g., use of evidence-based interventions) to influence student trajectories and thus the accuracy of risk predictions. Thus, a secondary goal of the present study is to examine potential practices that may remedy issues related to variations in screening performance. Specifically, heterogeneity in screening accuracy across both concurrent and predictive screening frameworks will be contrasted, wherein ASPENS either predicts performance on a concurrently administered criterion or after a several month delay.

II: METHOD

Design

This study represents a secondary analysis of data from a four-year, large-scale randomized control trial (RCT) conducted to evaluate the efficacy of the Fusion Math Intervention. The research design utilized a partially nested randomized controlled trial (RCT), with students randomly assigned to one of three treatment conditions using classroom-level randomization blocks. Students were assigned to receive Fusion Math in two of the conditions, which were distinguished by a student-teacher ratio of either 2:1 (2-Fusion) or 5:1 (5-Fusion). The third condition represented a business-as-usual (BAU) or no-treatment control condition. During their participation, all students continued to receive their district's core mathematics instruction.

Participants

Schools were recruited from Oregon and Massachusetts for two-years of participation (i.e., two cohorts) during the 2016-2017 and 2017-2018 school years. The sample was comprised of 26 schools across six districts. Across sites, 2,304 students were screened using the ASPENS, with 1,455 found eligible for participation in the randomized control trial. Eligibility was determined based on ASPENS composite scores. Specifically, students were considered eligible if their score fell in the Strategic or Intensive ranges based on ASPENS winter benchmarks. Among the 980 students that participated in the RCT, 291 were assigned to the BAU group, 192 to the 2-Fusion group, and 485 to the 5-Fusion group.

Student-level demographic data indicated that 55.78% of the students were female, 15.78% had special education status, and 15.02% had English learner status. The racial and ethnic distribution of the analytic sample included 0.43% American Indian or Alaska Native, 2.92% Asian, 3.46% Black or African American, 27.24% Hispanic or Latino, 0.65% Native Hawaiian or Other Pacific Islander, 8.00% reporting two or more races, and 57.30% White.

Instructional Conditions

Fusion Math

Fusion Math is a Tier 2 intervention intended to target whole number concepts and skills in Grade 1. It is comprised of 60 scripted lessons designed for a small group setting. Fusion Math was designed using principles of explicit and systematic instruction, incorporating such features as carefully scaffolded examples and practice items, frequent opportunities for student response, immediate teacher feedback, as well as strategic and structured review (Clarke et al., 2022). Empirical evidence supports the effectiveness of Fusion Math in improving proximal measures of math achievement such as fluency, problem-solving skills, and conceptual understanding (Cary et al., 2017; Clarke et al., 2014).

Importantly, as a Tier 2 intervention, Fusion Math is intended to be used in concert with high-quality whole-class instruction, not as a substitute. During the efficacy trial, the Fusion intervention was delivered outside of core instruction for five days per week across approximately 12 weeks. Thus, all condition groups received the same Tier 1 instruction; albeit instruction was inherently nested by classroom.

Measures

While the original study employed many additional measures, for the current study only two were utilized. Assessing Student Proficiency in Early Number Sense (ASPENS) was used as the screening measure, while Test of Early Mathematics Ability – 3rd Edition (TEMA-3) was used as the criterion measure.

ASPENS

ASPENS (Clarke et al., 2011) is a CBM screening assessment used to assess mathematical proficiency in kindergarten and first grade. It is composed of five measures: 1) Numerical Identification, 2) Magnitude Comparison, 3) Missing Number, and 4) Basic Arithmetic Facts and Base 10; with only three measures administered in a particular grade. In Grade 1, the focus of this study, only the Magnitude Comparison, Missing Number, and Basic Arithmetic Facts and Base 10 are administered. All ASPENS subtests are timed, with administration times ranging from one to two minutes depending on the subtest. Students' final scores represent the number of correct items within the elapsed time.

Delayed test-retest reliabilities for the ASPENS range from .76 to .85 in kindergarten and 0.77 to .87 in Grade 1. Concurrent validity coefficients relative to the mathematics subtest of the TerraNova 3rd Edition were .58 in kindergarten and .63 in Grade 1. Predictive validity coefficients of BOY ASPENS scores relative to EOY TerraNova scores were .53 in kindergarten and .57 in Grade 1.

As reported on the NCII Academic Screening Tool Chart (NCII, n.d.), the ASPENS cut-scores were identified to predict performance below the 15th percentile on mathematics subtest of the TerraNova – 3rd Edition. The authors indicate that thresholds were selected that were closest to achieving a sensitivity of .90. Sensitivity and specificity values for ASPENS cut scores are not reported on the NCII website as of June 2022, though AUC values exceed .80 at all times of year in Grade 1.

ASPENS includes three cut-scores, corresponding to the beginning-of-year (BOY), middle-of-year (MOY), and end-of-year (EOY). Because administration dates for cohorts varied such that they occurred in closer proximity to either the BOY or MOY screening periods, cut-scores for risk for differentially applied by cohort, based on the closest screening period to their median assessment date in their respective administration window.

TEMA-3

The TEMA-3 (Ginsburg & Baroody, 2003) is a standardized, individually administered, norm-referenced assessment designed to evaluate early mathematical skills in children between the ages of 3 and 8. This assessment is structured to evaluate a child's mathematical abilities across six essential domains: (1) numbering skills, encompassing counting and numeral recognition; (2) number-comparison facility, assessing the ability to compare and order numbers; (3) numeral literacy, focusing on reading and writing numerals; (4) mastery of number facts, including fluency in basic arithmetic; (5) calculation skills, such as adding, subtracting, multiplying, and dividing whole numbers; and (6) understanding of concepts, such as measurement, time, and geometric shapes. Test-restest reliability for the TEMA-3 ranges from .82 to .93 and alternate-form reliability is .97.

Student percentile ranks from the TEMA-3 were dichotomized for the purpose of defining math difficulty as a reference criterion. Specifically, student scores falling below the 20th percentile rank were considered cases for math difficulty and scores exceeding the 20th percentile will be considered non-cases. The 20th percentile was selected to approximate the definition of risk used to identify cut scores for the ASPENS (i.e., 15th percentile).

Importantly, the TEMA-3 relies on age-based norms rather than grade-based norms for its percentile ranking. Since participant age was only captured at the time of initial screening, rather than at the follow-up TEMA-3 assessment that is used in this study, student age was approximated by adding the number of elapsed days between each participants starting age and the median date of their respective cohort’s assessment window.

Analyses

Research Question 1

To answer Research Question 1, a generalized linear mixed model (GLMM) was specified investigate the effects of the 2:1 Fusion (2-Fusion) and 5:1 Fusion (5-Fusion) interventions on the ability of ASPENS to accurately identify cases and non-cases of math difficulty on the TEMA-3. Here, the outcome was represented as an accurate screening identification, or true negatives for non-cases or true positives for cases. The model is an extension of the models proposed by Riley et al. (2008), by focusing on the impact of group interventions on screening accuracy while accounting for within-school and across-school effects. Model 1 and Model 2 were first compared to determine the most suitable random effects structure. Following this comparison, Models 3 and 4 were compared to identify the most appropriate model for addressing the research question. The models were specified as follows:

Model 1

$$y_{ij} \sim Bernoulli\left(p_{ij}\right) \tag{1}$$

$$logit\left(p_{ij}\right) = \alpha_j \tag{2}$$

$$\alpha_j \sim N(0,\ \sigma_\alpha^2) \tag{3}$$

Model 2

$$y_{ij} \sim Bernoulli\left(p_{ij}\right) \tag{4}$$

$$logit\left(p_{ij}\right) = \alpha_j \ + \beta_{1j}\left(Sensitivity_{ij}\right) \tag{5}$$

$$\begin{pmatrix} \alpha_j \\ \beta_{1j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma\ \right]\ \Sigma = \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_{\beta_1} \\ \rho\sigma_\alpha\sigma_{\beta_1} & \sigma_{\beta_1}^2 \end{pmatrix} \tag{6}$$

Model 3

$$y_{ij} \sim Bernoulli(p_{ij}) \tag{7}$$

$$logit(p_{ij}) = \alpha_j + \beta_{1j}\,(Sensitivity_{ij}) + \beta_2\,(2\text{-}Fusion_{ij}) + \beta_3\,(5\text{-}Fusion_{ij}) + \beta_4\,(Sensitivity_{ij} \times 2\text{-}Fusion_{ij}) + \beta_5\,(Sensitivity_{ij} \times 5\text{-}Fusion_{ij}) \tag{8}$$

$$\begin{pmatrix} \alpha_j \\ \beta_{1j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma\ \right] \ \Sigma = \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_{\beta_1} \\ \rho\sigma_\alpha\sigma_{\beta_1} & \sigma_{\beta_1}^2 \end{pmatrix} \tag{9}$$

Model 4

$$y_{ij} \sim Bernoulli(p_{ij}) \tag{10}$$

$$logit(p_{ij}) = \alpha_j + \beta_{1j}\,(Sensitivity_{ij}) + \beta_2\,(2\text{-}Fusion_{ij}) + \beta_3\,(5\text{-}Fusion_{ij}) + \beta_4\,(Sensitivity_{ij} \times 2\text{-}Fusion_{ij}) + \beta_5\,(Sensitivity_{ij} \times 5\text{-}Fusion_{ij}) + \beta_6\,(Mean\_2\text{-}Fusion_{ij}) + \beta_7\,(Mean\_5\text{-}Fusion_{ij}) \tag{11}$$

$$\begin{pmatrix} \alpha_j \\ \beta_{1j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma\ \right] \ \Sigma = \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_{\beta_1} \\ \rho\sigma_\alpha\sigma_{\beta_1} & \sigma_{\beta_1}^2 \end{pmatrix} \tag{12}$$

Model 1 is described in Equations 1-3. The response variable y_{ij} represents the binary status of being accurately identified for student *i* in school *j*, using a Bernoulli distribution. The Bernoulli distribution is used to model binary data, with the probability of an accurate identification denoted by p_{ij}. In Model 1, α_j represents the random intercept for school j, accounting for the between-school variability in accurate identification. The random intercept for Model 1 is unique from the other models in that it represents the log-odds of accurate identification for both cases and non-cases, with no distinction for log-sensitivity and log-specificity.

Model 2, described in Equations 4-6, includes a fixed effect for the sensitivity group (i.e., cases; β_{1j}) to distinguish the probability of an accurate identification among cases and non-cases. By including this as a fixed effect with a random slope, the random intercept (α_j)

represents log-odds of accurate identification for non-cases, and the intercept for cases becomes represented by (α_j + β_{1j}). In other words, the compounded effect of the random intercept and the fixed effect for the sensitivity group captures the log-odds of accurate identification among cases, allowing for separate analyses of screening accuracy for cases and non-cases while considering school-level effects. Equation 7 describes the variance-covariance specification associated with the model. The random effects of μ_{1i} and μ_{0i} are specified such that that logit(p_{1i}) and logit(p_{0i}) are normally distributed around a mean logit-sensitivity and logit-specificity of β_1 and β_0 with a between-school variance of τ_1^2 and τ_0^2, respectively (Riley et al., 2008). Between-school correlation is represented using ρ, which accounts for the expected negative correlation between sensitivity and specificity.

Model 3 (Equations 7-9) and Model 4 (Equations 10-12) introduce additional fixed effects for treatment condition and are intended to answer the research question by modeling the effects of 2-Fusion and 5-Fusion on accurate identification for cases and non-cases. The model includes centered variables for 2-Fusion (β_2) and 5-Fusion (β_3) interventions, calculated as the difference between each student's individual group status and the average group status within their school. Prior to centering, students' group status was represented using dummy-coded indicator variables for each intervention group, which take the value of 1 if the student is in a particular intervention (either 2-Fusion or 5-Fusion) and 0 otherwise, indicating that they are in the BAU group. Students' group status was centered within schools to represent the difference between each student's individual group status and the average intervention participation within their school. This approach enables the model to account for within-school variations in intervention participation rates. Additionally, interaction terms for sensitivity are included for both 2-Fusion (β_4) and 5-Fusion (β_5) to distinguish the moderating effects of the treatment conditions for cases and non-cases.

Lastly, Model 4 includes fixed effects for the school-level means for 2-Fusion (β_6) and 5-Fusion (β_7) to control for average participation levels in the small and large group Fusion interventions. These variables help account for between-school variations in participation rates and enable a more precise estimate of intervention effects.

To convert model results to more interpretable values, marginal effects of Fusion intervention conditions for cases and non-cases were translated to sensitivity and specificity values on their traditional scale. Exponentiated effect estimates (i.e., odds ratios) were used to calculate the predicted probabilities of accurately identifying cases and non-cases for all possible combinations of sensitivity, 2-Fusion, and 5-Fusion, within the context of the research design These values were adjusted for school-level variations, or random effects of school in the GLMM model.

Research Question 1 Hypotheses. It was hypothesized that the accuracy of ASPENS in identifying non-cases would be systematically lower among students assigned to Fusion Math compared to a BAU condition. That is, it was expected that many students identified as cases using ASPENS at BOY would indeed have demonstrated math difficulties at that time, but that Fusion Math will effectively disrupt these predictions such that these students will present as FPs at EOY. Furthermore, these differences are expected to be greatest in the 2-Fusion condition, given that it represents more intense instructional support and would be expected to be more disruptive to forecasted trajectories. However, because Fusion Math does not have as significant implications for students that are already above the defined threshold for math difficulty, it is not expected to produce significant differences in sensitivity. Variations in screening accuracy among cases would only occur if students that do not present math difficulties earlier in the year fail to make sufficient progress such that they constitute cases and are rendered as false negatives at the end of year. If Fusion is an efficacious intervention, its effects on sensitivity are contingent on the extent that students above the ASPENS cut score do not make adequate progress in the BAU conditions. While variations in sensitivity are feasible between the treatment conditions, this was hypothesized to not reach statistical significance.

Research Question 2

To determine if treatment effects and variability in screening accuracy are diminished when ASPENS is administered concurrently with the TEMA-3, Models 3 and 4 will be replicated using EOY administrations of the ASPENS to predict TEMA-3 performance. That is, identical specifications of these models will be used with EOY TEMA-3 performance as the criterion, with EOY ASPENS scores used in substitute of BOY ASPENS. Notably, this will implicate a different cut score, given that the ASPENS cut scores vary depending on time of year. However, this is not a threat to the aim of the study, as the primary subject of interest will be the magnitude of observed effects on each cut score's sensitivity and specificity, rather than the sensitivity and specificity values themselves.

Research Question 2 Hypotheses. When comparing variability in the screening accuracy of ASPENS administrations conducted predictively or concurrently with the TEMA-3, it is hypothesized that the effects of 2-Fusion and 5-Fusion will be rendered non-significant in the concurrent analysis. Because Fusion Math will have preceded the administration of ASPENS in the concurrent analysis, students screening results are not expected to be vulnerable to the same instructional influences as the BOY ASPENS administration. Furthermore, variance is broadly expected to decrease in the random effects of all concurrent administration models due to other unmeasured school-level variables that will not have an opportunity to produce positive or negative risk migrations.

III: RESULTS

Descriptive statistics are presented in Table 4, summarizing the ASPENS pre-test and post-test scores, as well as the TEMA-3 scores for the three treatment groups (i.e., BAU, 2-Fusion, and 5-Fusion). The table displays the means, standard deviation, and sample sizes for the predictive and concurrent samples at both the individual and school levels. Additionally, classification counts and accompanying disaggregated means for ASPENS scores are included in Table 5.

Based on both school-level and participant-level descriptive statistics, ASPENS pre-test scores were generally comparable across the three conditions. Students assigned to the BAU condition demonstrated the highest mean score of 21.58 (*SD* = 11.52), followed by 5-Fusion at 20.62, and 2-Fusion with a mean score of 20.09. At post-test, 2-Fusion students had the highest ASPENS mean score of 47.23 (*SD* = 17.64), followed by 5-Fusion with a mean score of 43.77 (*SD* = 16.89), and the BAU condition with a mean score of 41.98. TEMA-3 scores demonstrated a similar pattern at posttest, with the BAU group demonstrating a mean score of 40.93, the 5-Fusion group with a mean of 41.62, and the 2-Fusion group showing the highest mean score of 42.45.

Table 3

Descriptive statistics for TEMA-3 and ASPENS scores in predictive and concurrent samples.

	Predictive Sample						Concurrent Sample					
			ASPENS		TEMA-3				ASPENS		TEMA-3	
Condition	*N*	Age (*M*)	*M*	*SD*	*M*	*SD*	*N*	Age (*M*)	*M*	*SD*	*M*	*SD*
						Students						
Control	255	6.67	21.58	11.52	40.95	8.17	259	6.67	41.98	18.35	40.93	8.19
2-Fusion	173	6.69	20.09	11.97	42.45	8.08	174	6.69	47.23	17.64	42.45	8.08
5-Fusion	430	6.67	20.62	11.16	41.64	8.17	431	6.67	43.77	16.89	41.62	8.17
						Schools						
Control	26	6.65	20.88	5.47	40.55	3.69	26	6.65	42.12	7.57	40.55	3.69
2-Fusion	26	6.68	20.22	6.21	43.00	4.24	26	6.68	47.26	9.57	43.00	4.24
5-Fusion	26	6.65	20.53	4.52	41.95	3.49	26	6.65	44.68	6.18	41.93	3.50

Note. *SD* = Standard Deviation; ASPENS = Assessing Student Proficiency in Early Number Sense. TEMA-3 = Test of Early Mathematics Ability – 3rd Edition. Age (*M*) = Age at beginning-of-year screening.

Table 4.

Classification Counts and ASPENS Means by Condition in Predictive and Concurrent Analyses

	Predictive Sample		Concurrent Sample	
Screening Result	Non-Cases *n (M)*	Cases *n* (*M*)	Non-Cases *n* (*M*)	Cases *n* (*M*)
Control				
Negative	TN = 133 (28.53)	FN = 31 (26.00)	TN = 126 (55.71)	FN = 17 (49.76)
Positive	FP = 43 (14.16)	TP = 53 (7.09)	FP = 49 (31.84)	TP = 67 (21.60)
5-Fusion				
Negative	TN = 208 (27.71)	FN = 52 (24.79)	TN = 210 (55.14)	FN = 47 (52.91)
Positive	FP = 90 (13.44)	TP = 82 (7.56)	FP = 87 (31.70)	TP = 87 (23.48)
2-Fusion				
Negative	TN = 84 (28.85)	FN = 13 (22.46)	TN = 96 (58.67)	FN = 8 (48.38)
Positive	FP = 47 (11.77)	TP = 30 (7.80)	FP = 35 (32.46)	TP = 35 (25.40)

Note. TN = True negative; FN = False negative; FP = False positive; TP = True Positive.

Research Question 1

Model Selection

Initially, Model 1 and Model 2 were compared to determine the suitable random effects structure. Model goodness of fit was evaluated by examining AIC, BIC, and R^2 values. Fit statistics indicated Model 2 was a better fit to the data, as evidenced by lower AIC and BIC, as well as higher marginal and conditional R^2 values compared to Model 1. Next, Models 3 and 4 were compared to determine the most appropriate model for addressing the research question. Specifically, Model 4 tested whether controlling for school-level intervention participation rates translated to improvements in overall model fit. However, Model 3 exhibited superior performance, with a lower AIC value and higher R^2 values compared to Model 4. Consequently, Model 3 was selected for further analysis and interpretation. Results for all models are summarized in Table 5.

Research Question 1 Results

For Model 3, the between-school variance for the random intercept corresponding to log-specificity (non-cases) was estimated at 0.54, with a standard deviation (SD) of 0.73. Between-school variance for the random slope of sensitivity was estimated at 2.10, with a standard deviation of 1.45. Random effect estimates suggest substantial variability in ASPENS screening accuracy across schools, with an intraclass correlation coefficient (ICC) of .14. The correlation between the random intercept and slope was -1.0.

Table 5.

Generalized linear mixed model results for predictive sample

	Model 1			Model 2			Model 3			Model 4		
	Coef.	SE	*p*	Coef.	SE	*p*	Coef.	SE	*p*	Coef.	SE	*p*
Fixed Effects												
Intercept, α	0.87	0.10	<.001	1.07	0.18	<.001	1.08	0.18	<.001	0.98	0.84	.24
Sensitivity, β_1	-0.32	0.16	.04	-0.68	0.34	.04	-0.69	0.34	.04	-0.62	0.35	.08
2-Fusion, β_2							-0.62	0.27	.02	-0.62	0.27	.02
5-Fusion, β_3							-0.29	0.23	.19	-0.29	0.23	.19
Sensitivity x 2-Fusion, β_4							0.89	0.52	.09	0.89	0.52	.09
Sensitivity x 5-Fusion, β_5							0.14	0.39	.73	0.13	0.39	.73
Mean 2-Fusion, β_6										0.91	1.63	.57
Mean 5-Fusion, β_7										-0.21	1.29	.87
Random Effects	Var.	SD		Var.	SD		Var.	SD		Var.	SD	
Intercept, α_j	0.02	0.14		0.54	0.73		0.55	0.74		0.53	0.73	
Sensitivity, β_{1j}				2.10	1.45		2.14	1.46		2.06	1.44	
Model Fit	AIC	BIC	R_c^2/R_m^2	AIC	BIC	R_c^2/R_m^2	AIC	BIC	R_c^2/R_m^2	AIC	BIC	R_c^2/R_m^2
	1085.72	1100.01	.01/.01	1045.43	1069.25	.16/.02	1047.02	1089.69	.17/.04	1050.70	1102.80	.17/.04

Note. Coef. = Coefficient; SE = Standard Error; Var. = Variance; SD = Standard Deviation, AIC = Akaike Information Criteria; BIC = Bayesian Information Criteria; R_c^2 = Conditional R-Squared; R_m^2= Marginal R-Squared

As predicted, the findings demonstrated that 2-Fusion had a significant negative effect on screening accuracy among non-cases (Estimate = -0.62, SE = 0.27, *p* = 0.02), suggesting that non-cases in this treatment condition were significantly less likely to be accurately differentiated from cases. The standardized effect size for 2-Fusion was -0.14 (95% CI [-0.32, 0.04]). Similarly, 5-Fusion demonstrated a negative effect on screening accuracy among non-cases; however, this was not statistically significant (Estimate = -0.29, SE = 0.23, *p* = 0.19), with a standardized effect size of -0.12 (95% CI [-0.30, 0.05]).

For log-sensitivity, the model revealed significantly lower screening accuracy compared to log-specificity (Estimate = -0.69, SE = 0.34, *p* = 0.04). The standardized effect size for sensitivity indicator variable was -0.32 (95% CI [-0.62, -0.01]). Consistent with the research hypotheses, the interaction term between 2-Fusion and log-sensitivity displayed a positive effect on screening accuracy (Estimate = 0.89, SE = 0.52, *p* = 0.09); however, this did not reach statistical significance at an alpha threshold of .05. The standardized effect size for this interaction term was 0.16, with a 95% confidence interval of [-0.02, 0.35]. Similar to 2-Fusion, the interaction term between 5-Fusion and the sensitivity indicator variable was positive but non-significant (Estimate = 0.14, SE = 0.39, *p* = 0.73). The standardized effect size for 5-Fusion's interaction term was 0.03 (95% CI [-0.14, 0.20]).

Model-predicted sensitivity and specificity values across the different conditions were generated using marginal effects to provide more interpretable values in describing ASPENS screening accuracy. For students in the BAU conditions, the predicted ASPENS sensitivity value was 0.59 (95% CI [0.42, 0.74]) and the predicted specificity was 0.81 (95% CI [0.72, 0.88]). For students assigned to 2-Fusion, the predicted sensitivity showed a moderate increase over BAU at 0.64 (95% CI [0.51, 0.75]), with predicted specificity significantly lower at 0.73 (95% CI [0.64, 0.80]). Lastly, for students assigned to the 5-Fusion condition, the predicted sensitivity of ASPENS was similar to the BAU condition at 0.55 (95% CI [0.42, 0.67]), with predicted specificity between that of 2-Fusion and BAU conditions at 0.77 (95% CI [0.68, 0.83]). Predicted sensitivity and specificity values, including standardized coefficients for fixed effects, are summarized in Table 6.

Research Question 2

Model Selection

Results from the concurrent models in this study indicated that no single model outperformed others across goodness of fit statistics. For instance, Model 1 excelled based on BIC, Model 2 exhibited the lowest AIC value, and Model 3 demonstrated the greatest R^2 values (conditional and marginal). Though some interaction parameters suggested potential relationships between treatment conditions and sensitivity, all predictors were found to be non-significant. Nonetheless, to properly contrast results with the predictive model results, Model 3 was selected for further elaboration. Results for all concurrent models are summarized in Table 6.

Research Question 2 Results

As illustrated in Table 6, the between-school variance for Model 3's random intercept corresponding to log-specificity (non-cases) was estimated at 0.19, with a standard deviation of 0.44. Between-school variance for the random slope associated with sensitivity (cases) was estimated at 0.55, with a standard deviation of 0.74. Random effects were much smaller compared to the predictive model, as evidenced by an ICC of .04 compared to .14. Similar to the predictive model, the correlation between the random intercept and random slope was -1.0.

The fixed effects for the concurrent model results indicate that 2-Fusion, yielded a non-significant effect on ASPENS ability to accurately detect non-cases (Estimate = 0.08, *p* = .76). Similarly, the 5-Fusion demonstrated no significant impact on screening accuracy of non-cases. (Estimate = -0.02, *p* = .91).

Among cases, the results suggest no significant difference in log-sensitivity compared to log-specificity (Estimate = 0.00, *p* = .99). When examining interaction terms, the interaction between 2-Fusion and the sensitivity indicator variable indicated that 2-Fusion effects were not significantly different among cases and non-cases (Estimate = 0.03, *p* = .96). The interaction between 5-Fusion and the sensitivity indicator variable was negative and most divergent from the magnitude of other parameter estimate but did meet the alpha threshold of .05 (Estimate = -0.71, *p* = .08). In summary, no fixed effects in the concurrent model were found to be significant at an alpha threshold of .05.

To capture the functional variations in sensitivity and specificity, marginal effects were also calculated using exponentiated effect estimates and are included in Table 7. In the BAU condition, the predicted sensitivity from the concurrent model was .79 (95% CI: [.65, .88]), while the predicted specificity was .72 (95% CI: [.63, .80]). Among students assigned to 2-Fusion, the predicted sensitivity increased to .80 (95% CI: [.51, .75]) and the predicted specificity increased to .74 (95% CI: [.67, .79]). For students in the 5-Fusion condition, the predicted sensitivity was .64 (95% CI: [.53, .74]), with the predicted specificity at .72 (95% CI: [.65, .78]).

Table 6.

Generalized linear mixed model results for concurrent sample

	Model 1			Model 2			Model 3			Model 4		
	Coef.	SE	*p*	Coef.	SE	*p*	Coef.	SE	*p*	Coef.	SE	*p*
Fixed Effects												
Intercept, α	0.93	0.09	<0.001	0.98	0.44	<0.001	0.98	0.13	<0.001	1.59	0.73	.03
Sensitivity, β_1	.04	0.16	.89	-0.02	0.22	.91	0.00	0.23	.98	0.05	0.23	.83
2-Fusion, β_2							0.08	0.27	.76	0.08	0.27	.76
5-Fusion, β_3							-0.02	0.22	.91	-0.02	0.22	.91
Sensitivity x 2-Fusion, β_4							0.03	0.56	.96	0.04	0.57	.83
Sensitivity x 5-Fusion, β_5							-0.71	0.41	.08	-0.71	0.41	.94
Mean 2-Fusion, β_6										0.17	1.35	.90
Mean 5-Fusion, β_7										-1.32	1.12	.24
Random Effects	Var.	SD		Var.	SD		Var.	SD		Var.	SD	
Intercept, α_j	0.01	0.12		0.19	0.44		0.19	0.44		0.17	0.41	
Sensitivity, β_{1j}				0.53	0.72		0.55	0.74		0.45	0.67	
Model Fit	AIC	BIC	R_c^2/R_m^2	AIC	BIC	R_c^2/R_m^2	AIC	BIC	R_c^2/R_m^2	AIC	BIC	R_c^2/R_m^2
	1032.43	1046.71	.00/.00	1029.75	1053.49	.05/.00	1030.79	1073.43	.06/.01	1033.14	1085.20	.06/.01

Note. Coef. = Coefficient; SE = Standard Error; Var. = Variance; SD = Standard Deviation, AIC = Akaike Information Criteria; BIC = Bayesian Information Criteria; R_c^2 = Conditional R-Squared; R_m^2= Marginal R-Squared

Table 7.

Summary of effect sizes and predicted sensitivity and specificity values

	Predictive Model		Concurrent Model	
Parameter	Std. Coef.	95% CI	Std. Coef.	95% CI
Sensitivity	-0.32	[-0.62, -0.01]	0.00	[-0.20, 0.21]
2-Fusion	-0.14	[-0.32, 0.04]	0.04	[-0.15, 0.22]
5-Fusion	-0.12	[-0.30, 0.05]	-0.12	[-0.29, 0.06]
Sensitivity x 2-Fusion	0.16	[-0.02, 0.35]	0.01	[-0.19, 0.20]
Sensitivity x 5-Fusion	0.03	[-0.14, 0.20]	-0.16	[-0.34, 0.02]
Condition	Sensitivity	Specificity	Sensitivity	Specificity
BAU	.59 [.42, .74]	.81 [.72, .88]	.79 [.65, .88]	.72 [.63, .80]
2-Fusion	.64 [.51, .75]	.73 [.64, .80]	.80 [.71, .87]	.74 [.67, .79]
5-Fusion	.55 [.42, .67]	.77 [.68, .83]	.64 [.53, .74]	.72 [.65, .78]

Note. Std. Coef. = Standardized coefficient.

IV: DISCUSSION

Schools are inherently diverse contexts due to various school-level factors, such as instructional methods, resources, teacher-student ratios, among other characteristics. Moreover, educational environments reflect a rich ecological system of influences (Bronfenbrenner, 1977), all of which may exhibit impacts on students' academic performance and growth during the school year. Despite these differences, universal cut-scores are often applied across schools based on presumed generalizability of their performance, potentially leading to an inaccurate assessment of students' needs and the inefficient allocation of resources should this notion of generalizability be erroneous.

Conventionally, academic screening tools are designed to predict student outcomes over time, such as in the case of a BOY screening tool predicting EOY performance. However, this approach introduces an opportunity for influence from the treatment paradox – a phenomenon that arises when lag time occurs between the administration of the screener and criterion measure. During this lag time, interventions are commonly conducted on the basis of screening results that presumably alter students' outcomes. In such cases, reflective evaluations of the diagnostic accuracy of screening measures can result in an over- or underestimation of a screener's effectiveness, because the intervention is intended to alter students' performance trajectories and can thereby skew the observed results of the screening assessment.

The purpose of this study was to conduct a more thorough investigation of this phenomenon in the context of early numeracy CBM screening accuracy across various instructional contexts within a RCT study for the Fusion Math intervention. This RCT study represented a context in which instruction was manipulated through random assignment, and different lag times between the screener and criterion measure could be compared. The present study's secondary analysis of the RCT results aimed to identify vulnerabilities and and investigate potential areas for improvement in universal screening measures and their accompanying cut-scores for MTSS decision-making.

Research Question 1: Heterogeneity in Predictive Screening Accuracy

The results of this study revealed that ASPENS screening accuracy differed between the three treatment groups (i.e., BAU, 2-Fusion, and 5-Fusion), as evidenced by patterns in descriptive statistics, results from the GLMM, and model-estimated sensitivity and specificity values. Despite comparable pre-test scores across the three conditions, 2-Fusion students had the highest mean scores on both ASPENS and TEMA-3, indicating greater gains compared to the 5-Fusion and BAU groups. These patterns are consistent with Clarke et al. (2022), who found that the greatest academic gains were found among students receiving 2-Fusion. Considering intervention-dependent changes in ASPENS screening accuracy are moderated by the intervention's efficacy, these results must be considered when reflecting on the findings of this study.

Specificity

Based on the results from the GLMM model, ASPENS screening accuracy for non-cases systematically varied across the Fusion Math Intervention conditions when compared to the BAU condition. However, based on Satterthwaite *p*-approximations, only 2-Fusion exhibited a significant effect ($p = .04$). Findings were consistent with the hypothesis that differences in screening accuracy would be more dramatic for non-cases (i.e., specificity), and that differences would be most evident within the 2-Fusion condition. In short, non-cases (i.e., scoring above the 20th percentile on the EOY TEMA-3) in the 2-Fusion condition were more likely to be falsely categorized as “at-risk” based on their BOY ASPENS screening assessment. Whether these students indeed constitute false positive errors or represent instances of positive risk migration will be explored in the subsequent section. Nonetheless, an agnostic review of ASPENS misclassification errors found that they translated to an estimated specificity value of .73 (95% CI [.64, .80]) for the 2-Fusion group compared to .81 (95% CI [.72, .88]) in the BAU condition. Estimated specificity for the 5-Fusion condition fell in between the 2-Fusion and BAU condition at .77 (95% CI [.64, .80]), suggesting that it exhibited a similar influence on the latent math abilities of students as 2-Fusion, but was not as disruptive to students’ performance trajectories to be meaningfully divergent from the BAU condition. However, were these values to be interpreted as part of an evaluation of ASPENS screening accuracy, such as for the National Center for Intensive Intervention’s Academic Screening Tools Chart (NCII, 2020), ASPENS cut-scores would be found to perform below NCII’s acceptable specificity standards of .80 in both Fusion groups yet would meet acceptable performance within the BAU condition.

Sensitivity

It was hypothesized that the Fusion intervention conditions would also demonstrate some influence on sensitivity, but differences would not be as prominent compared to specificity. Because risk migration errors in sensitivity result from students experiencing a decline in their skills relative to the normal distribution, the Fusion intervention was expected to serve as a protective factor against negative risk migration by maintaining or bolstering students’ skills and preventing normative declines. However, the scope to which this protective influence meaningfully moderates sensitivity values would be more dependent on the prevalence of negative risk migration among the BAU condition. For example, if negative risk migration is not prevalent in the BAU condition, this protective influence would be unable to translate to significant differences between the BAU and intervention groups in this sample, as there would not be enough instances for Fusion to “avert.” Nonetheless, the Fusion conditions were expected to produce slightly higher sensitivity values due to fewer instances of negative risk migration, but that these would not be in sufficient frequency among the BAU condition to translate to meaningful variations in sensitivity.

As hypothesized, results from the concurrent GLMM found that 2-Fusion and 5-Fusion showed a positive effect on ASPENS screening accuracy among cases. However, these values were not statistically significant based on Satterthwaite *p*-value approximations. Similar to the findings for specificity, 2-Fusion exhibited the greatest difference in screening accuracy compared to the BAU condition (Estimate = 0.89, SE = 0.52, *p* = 0.09).

In general, model-predicted sensitivity values demonstrated similar variability across conditions to specificity, but not necessarily as anticipated. The 5-Fusion condition demonstrated the lowest sensitivity at .55 (95% CI [.42, .67]) and 2-Fusion had the highest sensitivity value at .64 (95% CI [.51, .75]). Predicted sensitivity for the BAU condition was .59 (95% CI [.42, .74]). Interestingly, it was unexpected for both the variance in sensitivity values across conditions to be similar to specificity and for 5-Fusion to demonstrate the lowest predicted sensitivity value.

Notably, the classifications within a 2x2 confusion matrix that contribute to strongly dictate differences in sensitivity (i.e., false negatives) represent the smallest cell counts for each condition. Furthermore, standard errors of the model coefficients related to sensitivity for noticeably higher than for specificity. Thus, the absence of significant coefficient estimates at an alpha threshold of .05, larger standard errors of the estimates, and smaller cell counts suggest that these patterns should be interpreted with caution.

Evidence for Positive and Negative Risk Migration

The concepts of positive and negative risk migration may play a critical role in understanding variability in screening accuracy across instruction and intervention conditions, as the lag time in predictive models permits true changes in students' academic performance between the administrations of the screening and criterion assessment. The RCT design of the Fusion efficacy study, and the fact that intervention assignment occurred at the student level nested within schools, enabled the examination of the influence of quality and format of instruction for their role in this phenomenon. Indeed, both forms of risk migration can be influenced by factors such as the quality of core instruction, supplemental supports, curricular alignment, individual fluctuations in development or academic growth, and various other contextual variables. However, this study hypothesized that randomized assignment to the Fusion intervention or the BAU condition would most likely moderate occurrences of positive risk migration such that Fusion conditions would result in higher incidence of false positives, translating to observed variations in specificity. Variations in sensitivity were less expected due to the RCT study design in which students were exposed to the same core instruction and thus systematic variations in negative risk migration were less likely to be evident across the independent variables included in the model (i.e., Fusion Math vs. BAU).

To review the concepts put forth in this study, risk migration refers to transitions between categories (e.g., typical-achievement population, academic-difficulty population) that are not attributable to measurement error but instead result from environmental efforts to alter the category an individual belongs to. Naturally, these categories may be contrived, such as in the case of this study, wherein academic difficulty was defined as performance below the 20th percentile. It is important to recognize that some risk migration can be expected as an inherent consequence of dichotomizing continuums that are not stable over time, such distributions of academic performance. By adopting fixed cut-points for an unstable continuum, some individuals will naturally drift across cut-points over time with little true change in their ability. However, the underlying assumption made in the current interpretation is that students' position within the normative distribution has differentially shifted across conditions between the ASPENS and TEMA-3 administrations.

To further inform the tenability of this assumption, Table 4 helps to illuminate differences in performance among students who were misclassified by ASPENS. Among positive non-cases (FP) in the predictive sample, the 2-Fusion condition had a meaningfully lower mean ASPENS score (*M* = 11.77) compared to the BAU condition (*M* = 14.16) and the 5-Fusion condition (*M* = 13.44), as illustrated in Table 4. This lower mean value for the 2-Fusion group suggests that students within the 2-Fusion intervention experienced greater improvement in their math performance. In other words, lower mean scores on the ASPENS among non-cases indicates that students met the threshold for typical achievement (i.e., scoring above the 20th percentile on TEMA-3) despite lower pretest performance on the ASPENS. This pattern suggests further indication of positive risk migration. That is, students with lower ASPENS scores migrated from the at-risk population to the typical-achievement population ostensibly due to the intervention's influence, effectually broadening the feasibility of attaining typical achievement by EOY into lower distributional regions of the screening tool at BOY. Consequently, greater propensities for positive risk migration would theoretically possess an inverse association with the average scores of FPs on the screening assessment within the most intensive instructional group, as was observed in the data.

Additionally, when analyzing the negative cases (FN) in the predictive sample, the lower mean values observed in the 2-Fusion (*M* = 22.46) and 5-Fusion (*M* = 24.79) conditions, as compared to the BAU condition (*M* = 26.00), hint at the potential success of the 2-Fusion and 5-Fusion interventions in mitigating negative risk migration. Recall that negative risk migration refers to the phenomenon where students transition from the typical achievement population to the at-risk population. The observed lower mean values for FNs in the intervention conditions suggest that they may have helped students to maintain their academic performance by averting a decline in their performance relative to the BAU condition. Similar to the Will Rogers phenomenon, or stage migration bias (Howard, 2019), reclassification of individuals between groups can produce increases in mean values of both groups. In the context of this study, when students from the typical achievement population undergo negative risk migration into the at-risk population, the mean performance in the at-risk group can increase as it gains students who, although struggling, are likely performing better than its existing members. It remains unclear why sensitivity was lowest in the 5-Fusion condition despite adhering to this expected pattern in the descriptive data. However, the most likely explanation is that the suppressive effect of 5-Fusion on negative risk migration, while possibly suggested in the descriptive results, was not sufficient to manifest in sensitivity and overcome random sampling variability, as substantiated in the model results.

Lastly, it is important to note that while these descriptive results highlight patterns suggestive of the potential presence of risk migration, they only serve as symptoms or trends related to these phenomena, rather than direct evidence.

Research Question 2: Reductions in Heterogeneity in Concurrent Administrations

Research Question 2 investigated the extent to which heterogeneity in screening accuracy could be mitigated when ASPENS and TEMA-3 were administered concurrently, rather than at different times of the year (i.e., BOY vs. EOY). This analysis attempted to provide insights into the role of temporal factors in mediating the influence of measured and unmeasured variables on ASPENS screening accuracy. Premised on the notion that the temporal relationship between the screening and criterion assessment is an important contributor to variability in screening accuracy, it was hypothesized that any notable variations across treatment conditions observed in the predictive ASPENS administration would be nullified with concurrent administrations.

Indeed, the concurrent model demonstrated a noteworthy reduction in heterogeneity across schools and treatment conditions compared to the predictive model. The amount of variance in sensitivity and specificity fell from .14 for the predictive model to .05 for the concurrent model, suggesting less variability in screening accuracy was attributable to unobserved school-level effects. Differences in sensitivity and specificity were also less pronounced between treatment conditions in the concurrent assessment models. Results from the concurrent GLMM found the effects of 2-Fusion and 5-Fusion on screening accuracy were non-significant for both cases and non-cases, as assessed using Satterthwaite *p*-value approximations.

With concurrent administrations of the ASPENS and TEMA-3, specificity values exhibited only minor fluctuations between treatment conditions. Both the BAU condition and 5-Fusion produced specificity values of .72, with 2-Fusion demonstrating a modest improvement at .74 (95% CI [.67, .79]). Predicted sensitivity values were similarly consistent between 2-Fusion and BAU conditions, as indicated by sensitivities of .80 (95% CI [.51, .75]) and .79 (95% CI [.65, .88]), respectively. Notably, the predicted sensitivity of the 5-Fusion condition was lower at

.64 (95% CI [.53, .74]), though not significantly so. Nonetheless, in the absence of statistically significant model coefficients, these differences in sensitivity must be interpreted with caution.

Relevance to Field

Building on research that has reported poor generalizability of screening tool cut-scores across educational settings (Klingbeil et al., 2015, 2018, 2021, 2022; Hintze et al., 2003; Johnson et al., 2009), this study is one of the first to directly model the repercussions of poor generalizability and offer a more detailed explanation for one source of poor generalizability: heightened positive risk migration following the provision of more effective instruction. Contrary to historical assertions about the applicability of certain diagnostic accuracy indices in universal screening practices (Johnson et al., 2009; Petscher et al., 2011a; Van Norman et al., 2016; Vanderheyden, 2011), the current findings revealed that the specificity of an early numeracy screening tool covaried with instructional conditions after accounting for other school-level factors. Importantly, these findings likely extend to other assessments and highlight a vulnerability that is not theoretically unique to the screening tool studied here. Taken together, the outcomes of this study show that the concepts of positive and negative risk migration indeed hold explanatory power and relevance within real-world educational settings, where changes in students' abilities are subjected to influence by factors including the quality of core instruction, curricular alignment, provision of supplemental supports, individual fluctuations in development or academic growth, and a host of other contextual variables.

The predominance of prognostic screening frameworks, which often rely on EOY measures as criteria for all screening periods, permits such educational factors the opportunity to influence students’ trajectories, thereby introducing sample-specific biases which can be misinterpreted as screening tool error. Thus, it is critical to recognize the potential influences of positive and negative risk migration when interpreting screening assessment results and evaluating the accuracy of screening instruments from both psychometric and practical perspectives.

Observed inconsistencies in the accuracy of screening assessments when using publisher-recommended cut-scores across different school settings have prompted some calls for local validation in ensuring the efficacy of such tools (Keller-Margulis et al., 2008; Klingbeil et al., 2012, 2022; Nelson et al., 2017; Patton et al., 2014; Thomas & January, 2019). Indeed studies have shown that adhering to a single nationwide cut-score may lead to imbalances and inadequate screening accuracy in both mathematics and reading (Klingbeil et al., 2015, 2018, 2021; Hintze et al., 2003; Johnson et al., 2009). Thus, researchers and educators should thoroughly consider the use of local cut-scores tailored to the specific characteristics and needs of their student population. However, this study highlights one shortcoming that would remain unresolved. That is, the process of determining cut-scores within a prognostic screening framework still misattributes positive and negative risk migration to screening tool inaccuracies, even if the validation sample is better aligned with the local context for its intended use. For example, were the 2-Fusion sample to be utilized to identify a cut-score that minimizes screening error, conventional methods would retrospectively attempt to avert using scores that had previously identified students in the FP category as "at-risk" on the screening tool. Consequently, in the future, students that underwent positive risk migration after receiving 2-Fusion would likely not have been candidates for the intervention under a newly "calibrated" cut-score. Furthermore, students with similar scores in the future may indeed present as FNs, since the new selection process failed to previously account for the influence of positive risk migration, disallowing students of the supports necessary for them to exceed the threshold for academic difficulty by EOY. Theoretically, were these dynamics to be illustrated through simulated cut-score calibrations based on sensitivity and specificity values, with randomly varying intervention effects applied to students falling below the new cut-score, researchers would likely see patterns of reactive increases and decreases due to these misjudgments about the nature of screening errors.

Considerations to Improve Generalizability

As expected, use of a concurrent screening model resulted in less heterogeneity in screening accuracy and mitigated the treatment effects of the intervention observed in a predictive model, given that concurrent administrations precluded the theoretical ability for risk migration to occur. While the use of EOY screening tools is commonplace (Glover & Albers, 2007; Vanderheyden et al., 2018), most screening assessments remain anchored to an EOY criterion at all timepoints (Smolkowski & Cummings, 2015), and significant emphasis is placed on the use of screening tools for forecasting student outcomes (Ball & O'Connor, 2016; Chard et al., 2008; Petscher et al., 2011; Roehrig et al., 2008; Yeo, 2010). Until relatively recently, NCII required screening tools to implement a lag time of at least three months with their criterion measure as a prerequisite to their evaluation process (NCII, 2018), thus entrenching prognostic models into the validation history of existing tools. Going forward, a potential remedy to the issues raised by Research Question 1 is to reconsider the use of prognostic screening approaches and expand the use of concurrent models, which are often already implicitly in use for EOY screening assessments. By reducing lag time between the administration of the screener and criterion measure, decision-making may be less biased by sample characteristics that moderate growth over time, such as instructional quality.

The adoption of a concurrent screening model entails focusing more on evaluating students' status at a single moment in time, rather than predicting future status based on undefined assumptions about the instructional supports provided during that time. Thus, brief screening assessments' psychometric evidence could be anchored to more comprehensive criterion measures appropriate for students' developmental stage in math and reading. Interpretively, screening results would then describe students expected performance status were a more extensive assessment administered.

Admittedly, fully adopting a concurrent screening model presents challenges, as identifying appropriate definitions of proficiency at each screening period implicates more judgements and would need to be done thoughtfully. This process should aim to identify benchmarks that would be less reliant on predictive assumptions that do not generalize across educational settings. Alternatively, it may be feasible to explore methods within a prognostic framework that appropriately compensate for contextual factors, such as the use of correction factors or other analytic approaches that help ensure that cut-score selection is not unduly influenced by biases attributable to risk migration.

Importantly, advocating for a concurrent screening framework does not invalidate or discount the utility of predictive analyses, nor the process of forecasting student trajectories. However, these activities may be better relegated to other contexts such as research or systems-level evaluations, where precision may not be as consequential to the decision-making for individual students.

Limitations

This study has several limitations that should be considered when interpreting its findings. First, due to differences in BOY screening dates, the BOY and MOY ASPENS cut scores were applied to cohorts depending on assessment date. That is, as noted in the Methods section, some cohorts were assessed sufficiently late in the Fall that administration dates occurred much closer to the MOY assessment period than BOY period. As a result, cut scores for the most proximal screening period were applied to each cohort. Consequently, the predictive model does not characterize the screening accuracy of a single ASPENS cut-score, but rather represents a blend of the BOY and MOY cut-score.

In addition, the current study did not control for intervention fidelity, which could have led to variations in the delivery and quality of the Fusion implementation. Clarke et al., (2022) reported in their analyses of a subset of participating schools that 2-Fusion groups demonstrated greater total fidelity (*g* = 0.25) than 5-Fusion groups. Additionally, 2-Fusion groups were rated higher by trained observers on meeting instructional objectives, use of prescribed models, frequency of teaching activities, and adherence to scripting. Differences in implementation fidelity may have contributed to the greater differences in screening accuracy found among the 2-Fusion than the 5-Fusion group. If intervention fidelity data or other measures of instructional characteristics were available for all participating groups, it could offer clearer insight into differences between 2-Fusion and 5-Fusion effects and whether they are attributable to intensity as defined by group size, by fidelity, or both. Intervention fidelity was not explored as a moderator in the current study to leverage the maximum sample size because fidelity was not measured in all schools.

Lastly, the current study focused only on positive and negative risk migration only as related to early numeracy skills in a single study. Furthermore, this study investigated one definitional threshold for academic difficulty on a single criterion measure (i.e., TEMA-3). It is unclear how the use of different screening and criterion measures may have altered current findings. Thus, there is need for more extensive research in this area, both within the context of early mathematics screening and intervention, with other academic domains (e.g., reading, writing), and with multiple criterion measures.

Future Directions

Further investigation into heterogeneity in screening accuracy, including the influence of both student- and school-level factors, can help build a deeper understanding of how early screening assessments function across various settings. Such information is necessary not only for measure development and refinement, but to accurately communicate their practical utility to consumers (i.e., educators), and support informed decision-making as it relates to individual students, resource allocation, and systems-level evaluations.

In pursuit of these goals, the present study represents the first application of a generalized linear mixed model (GLMM) in educational research to jointly model sensitivity and specificity using participant-level data. This methodology was adapted from individual participant data meta-analytic (IPD-MA) techniques, which have become central in the synthesis of modern healthcare research and have been bolstered by the open-science movement (Macaskill et al., 2010; Riley et al., 2021). The use of IPD-MA has historically been challenging across research contexts, as it requires individual-level data from each unit in the meta-analysis. However, as demonstrated, the analytic methods can be leveraged effectively to account for the nesting structures encountered in educational research. The use of the analytic methods adopted in this study, as well as others from IPD-MA research such as hierarchal summary receiver operating characteristic (HSROC; (Harbord et al., 2008; Rutter & Gatsonis, 2001) models, offers the opportunity for educational researchers to explore many more research questions related to screening accuracy.

In particular, accounts of variability cut-score performance remain prevalent in the literature (e.g., Klingbeil et al., 2012, 2022; Hintze et al., 2003), but research exploring and accounting for this variability remains is limited. Opportunities for further investigation include applying HSROC models to examine heterogeneity in screening performance across all cut-scores, exploring the influence of other measurable school-level factors on cut-score performance such as the base rate of academic difficulties (i.e., spectrum effects), and investigating how patterns in heterogeneity manifest across different measurement tools. This information would be useful not only for refining the development of universal screening tools, but also establishing a better understanding of their appropriate uses and misuses in student and systems-level decision-making.

Implications for practice

The findings of this study suggests that evaluations of screening accuracy, such as what is currently depicted in the NCII screening tools chart (NCII, 2018) should be interpreted as estimates across a particular aggregated sample rather than a true summary of screening accuracy that is directly generalizable to consumers. That is, current validation procedures summarize classification accuracy within the context of a study sample, not at the student- or school-level where decisions are applied. These shortcomings primarily become a concern if accuracy indices express variability across student and school contexts, which was identified in this study when lag time occurs between the screening and criterion assessment.

Additionally, consumers may consider more carefully applying scrutiny to the normative representation of screening validation samples, as is done to establish other psychometrics such as percentile ranks. Notably, to generalize the accuracy of predictions, requests for more information about sample characteristics particularly regarding the school contexts, in addition to demographic characteristics of the student themselves. Nonetheless, there is some evidence for inconsistencies in predictive validity for early literacy tools across student-level demographic features (Hosp et al., 2011).

Lastly, educators are discouraged from establishing local cut scores without an accompanying process to review misclassifications and rule out positive and negative risk migration. For example, if schools want to pursue predictive cut scores, progress monitoring data could be reviewed to determine whether FPs were in fact instances of response to intervention. Characteristic features of positive risk migration would likely include: a) low performance at initial screening, b) responsive provision of supplemental instructional supports, c) and above average growth in progress monitoring data that coincided with the intervention's onset. Conversely, occurrences of negative risk migration would likely be more difficult to identify since progress monitoring would be unavailable due to being "missed" by the screener. However, educators could suspect negative risk migration to manifest where divergences in the quality of core instruction occur, such as across classrooms within a school or schools within a district. If there is a higher density of FNs in a particular classroom or school, it would be likely that lower rates of academic growth in that setting prompted occurrences of negative risk migration rather than suggesting a miscalibrated cut-score across the entire system.

If patterns in FPs or FNs do show evidence of risk migration, they should be considered as accurate predictive identifications when determining appropriate cut scores for a particular school or district because students migrated presumably due to response to intervention. Nonetheless, educators should still be aware of the close relationship that exists between screening predictions and instructional contexts, which may drift as practices change over time. Factors such as staffing, student demographics, curricula, and other school features will inevitably shift, altering the assumptions underlying any forecasts made about student trajectories and risk of academic difficulties in a particular setting.

Conclusion

The current study posited that commonly used indices of predictive accuracy for early academic screening instruments and their accompanying cut scores is more context-dependent than commonly thought. Two phenomena were defined as mechanisms for this variability – positive risk migration and negative risk migration. These terms were used to describe the concepts of students moving across an established threshold for academic difficulty due to instructional influences, whereupon students either move from the typical achievement population to the academic difficulty population due to inadequate growth (i.e., negative risk migration), or start the year in the academic difficulty population but demonstrate sufficient academic growth to enter the typical achievement population by the end of the year.

To highlight the practical implications of migration and its effects on predictive accuracy, this study conducted a secondary analysis of a math intervention RCT study with the hypothesis that its randomized assignment to intervention conditions would accentuate instances of positive risk migration, such that meaningful differences in specificity would manifest in predictive accuracy for the screening measure used. Indeed, systematic variation in specificity was observed as hypothesized among students assigned to the most intensive intervention condition compared to a BAU condition. Furthermore, these differences in specificity no longer manifested when the screening measure was administered concurrently with its criterion measure post-intervention, meaning that instruction was not manipulated during a time lag in between the administration of the screening and criterion assessments.

Consequently, researchers and educators are encouraged to exercise more caution when interpreting cut score performance relative to specific settings. Arguably, within the context of MTSS with evidence of response to intervention, whether experimental or not, it would be inappropriate to consider students who underwent positive or negative risk migration as misclassifications on the screening assessment because these students underwent a genuine change in status due to intervention.

Before the widespread adoption of MTSS systems, the phenomena of positive and negative risk migration may not have been as pronounced. However, educators' now ubiquitous and concerted efforts to intervene and alter predicted trajectories, through the use of screening data and evidence-based interventions, makes it increasingly untenable to ignore these as contextual factors when validating screening tools. Nonetheless, current practice in screening validation still considers such students as misclassifications and must continue to do so until procedures are established account for them. Thus, measure developers and consumers are urged to advocate for better specification of the instructional conditions under which screening accuracy is determined. Regardless, setting intervention effects aside, screening accuracy showed much greater variability across schools for predictive screening accuracy than for concurrent screening accuracy, suggesting the field may benefit from less confusion regarding false positives were a concurrent approach to cut-score selection to be adopted. More research is necessary to determine the measurable impact of positive and negative risk migration on screening accuracy variability for other assessment tools and academic domains, as well as how to appropriately account for these phenomena when setting predictive cut scores and evaluating them.

REFERENCES

Balu, R., Zhu, P., Doolittle, F., Schiller, E., Jenkins, J., & Gersten, R. (2015). Evaluation of Response to Intervention Practices for Elementary School Reading. NCEE 2016-4000. *National Center for Education Evaluation and Regional Assistance*.

Brenner, H., & Gefeller, O. (1997). Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Statistics in Medicine*, *16*(9), 981–991.

Bronfenbrenner, U. (1979). The Ecology of Human Development: Experiments by Nature and Design. Harvard University Press.

Catts, H. W., Nielsen, D. C., Bridges, M. S., Liu, Y. S., & Bontempo, D. E. (2015). Early Identification of Reading Disabilities Within an RTI Framework. Journal of Learning Disabilities, 48(3), 281–297. https://doi.org/10.1177/0022219413498115

Clarke, B., Gersten, R. M., Dimino, J., & Rolfhus, E. (2011). *Assessing student proficiency of number sense (ASPENS)*. Longmont, CO: Cambium Learning Group, Sopris Learning.

Clarke, B., Doabler, C. T., Sutherland, M., Kosty, D., Turtura, J., & Smolkowski, K. (2022). Examining the Impact of a First Grade Whole Number Intervention by Group Size. *Journal of Research on Educational Effectiveness*. https://doi.org/10.1080/19345747.2022.2093299

Clarke, B., Doabler, C. T., Cary, M. S., Kosty, D., Baker, S., Fien, H., & Smolkowski, K. (2014). Preliminary evaluation of a Tier 2 mathematics intervention for first-grade students: Using a theory of change to guide formative evaluation activities. School Psychology Review, 43(2), 160–177. https://doi.org/10.1080/02796015.2014.12087442

Clemens, N. H., Shapiro, E. S., & Thoemmes, F. (2011). Improving the efficacy of first grade reading screening: An investigation of word identification fluency with other early literacy indicators. *School Psychology Quarterly, 26*(2), 231-244. https://doi.org/10.1037/a0025173

Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, *115*(7), 928–935. https://doi.org/10.1161/CIRCULATIONAHA.106.672402

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional children*, *52*(3), 219-232.

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2004). Monitoring early reading development in first grade: Word identification fluency versus nonsense word fluency. *Exceptional Children, 71*(1), 7–21. https://doi.org/10.1177/001440290407100101

Gambino, B. (2006). Reflections on accuracy. *Journal of Gambling Studies*, *22*(4), 393–404. https://doi.org/10.1007/s10899-006-9025-5

Gambino, B. (2018). Test performance variation between settings and populations. *Journal of Gambling Studies*, *34*(4), 1085–1108. https://doi.org/10.1007/s10899-017-9728-9

Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J. R., & Witzel, B. (2009). Assisting students struggling with mathematics: Response to intervention (RtI) for elementary and middle schools. *IES National Center for Education Evaluation Practice Guide*.

Ginsburg, H., & Baroody, A. (2003). TEMA-3 examiners manual. *Austin, TX: Pro-Ed*.

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology*, *45*(2), 117–135. https://doi.org/10.1016/j.jsp.2006.05.005

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36.

Hintze, J. M., Ryan, A. L., & Stoner, G. (2003). Concurrent validity and diagnostic accuracy of the Dynamic Indicators of Basic Early Literacy Skills and the Comprehensive Test of Phonological Processing. *School Psychology Review, 32,* 541–556. https://doi.org/10.1080/02796015.2003.12086220

Howard, J. (2019). *Cognitive Errors and Diagnostic Mistakes*. Springer International Publishing. https://doi.org/10.1007/978-3-319-93224-8

January, S. A. A., & Klingbeil, D. A. (2020). Universal screening in grades K-2: A systematic review and meta-analysis of early reading curriculum-based measures. *Journal of School Psychology*, *82*, 103-122.

Jenkins, J. R. (2003, December). Candidate measures for screening at-risk students. Paper presented at the National Research Center on Learning Disabilities' Responsiveness-to-Intervention Symposium, Kansas City, MO.

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review, 36*(4), 582–600. https://doi.org/ 10.1080/02796015.2007.12087919

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How can we improve the accuracy of screening instruments?. *Learning Disabilities Research & Practice*, *24*(4), 174-185.

Kameenui, E. J., & Carmine, D. W. (Eds.). (1998). *Effective strategies for accommodating students with diverse learning and curriculum needs*. Columbus, OH: Merrill.

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, *42*(4), 448-457. https://doi.org/10.1080/02796015.2013.12087465

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

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

Klingbeil, D. A., McComas, J. J., Burns, M. K., & Helman, L. (2015). Comparison of predictive validity and diagnostic accuracy of screening measures of reading skills. *Psychology in the Schools*, *52*(5), 500-514.

Klingbeil, D. A., Nelson, P. M., van Norman, E. R., & Birr, C. (2017). Diagnostic Accuracy of Multivariate Universal Screening Procedures for Reading in Upper Elementary Grades. *Remedial and Special Education*, *38*(5), 308–320. https://doi.org/10.1177/0741932517697446

Klingbeil, D. A., Osman, D. J., Van Norman, E. R., Berry-Corie, K., Kim, J. S., Schmitt, M. C., & Latham, A. D. (2023). Universal Screening with aimswebPlus Reading in Middle School. *Reading & Writing Quarterly*, *39*(3), 192-211.

Klingbeil, D. A., van Norman, E. R., Nelson, P. M., & Birr, C. (2019). Interval likelihood ratios: Applications for gated screening in schools. *Journal of School Psychology*, *76*, 107–123. https://doi.org/10.1016/j.jsp.2019.07.016

Leeflang, M. M., Bossuyt, P. M., & Irwig, L. (2009). Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. *Journal of clinical epidemiology*, *62*(1), 5-12.

Leeflang, M. M. G., Rutjes, A. W. S., Reitsma, J. B., Hooft, L., & Bossuyt, P. M. M. (2013). Variation of a test's sensitivity and specificity with disease prevalence. *Cmaj*, *185*(11), 537–544. https://doi.org/10.1503/cmaj.121286

Macaskill, P., Gatsonis, C., Deeks, J., Harbord, R., & Takwoingi, Y. (2010). Cochrane handbook for systematic reviews of diagnostic test accuracy.

McCardle, P., Scarborough, H. S., & Catts, H. W. (2001). Predicting, explaining, and preventing children's reading difficulties. *Learning disabilities research & practice*, *16*(4), 230-239.

Mulherin, S. A., & Miller, W. C. (2002). Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Annals of Internal Medicine*, *137*, 598–602.

National Center on Response to Intervention. (n.d.). *Academic Screening Tools Chart*. Washington, DC: U.S. Department of Education, Office of Special Education Programs, National Center on Response to Intervention. Retrieved from https://charts.intensiveintervention.org/ascreening

National Center on Response to Intervention. (2020). *Academic Screening Tools Chart Rating Rubrics*. Washington, DC: U.S. Department of Education, Office of Special Education Programs, National Center on Response to Intervention. Retrieved from https://intensiveintervention.org/sites/default/files/NCII_AcademicScreening_RatingRubric_2020-06-30.pdf

Nelson, P. M., Van Norman, E. R., & VanDerHeyden, A. (2017). Reduce, reuse, recycle: The longitudinal value of local cut scores using state test data. *Journal of Psychoeducational Assessment*, *35*(7), 683–694. https://doi.org/10.1177/0734282916658567

Petscher, Y., Fien, H., Stanley, C., Gearin, B., Gaab, N., Fletcher, J.M., & Johnson, E. (2019). *Screening for Dyslexia.* Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education, Office of Special Education Programs, National Center on Improving Literacy. Retrieved from improvingliteracy.org.

Petscher, Y., Kim, Y.-S., & Foorman, B. R. (2011). The importance of predictive power in early screening assessments: Implications for placement in the response to intervention framework. *Assessment for Effective Intervention*, *36*(3), 158–166. https://doi.org/10.1177/1534508410396698

Ransohoff, D. F., & Feinstein, A. R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine*, *299*(17), 926-930.

Reschly, A. L., Busch, T. W., Betts, J., Deno,S. L., & Long, J. D. (2009). Curriculum- based measurement oral reading as an indicator of reading achievement: A meta- analysis of the correlational evidence. *Journal of School Psychology, 47*(6), 427- 469. https://doi.org/10.1016/j.jsp.2009.07.001

Riley, R. D., Dodd, S. R., Craig, J. v., Thompson, J. R., & Williamson, P. R. (2008). Meta-analysis of diagnostic test studies using individual patient data and aggregate data. *Statistics in Medicine*, *27*(29), 6111–6136. https://doi.org/10.1002/sim.3441

Riley, R. D., Stewart, L. A., & Tierney, J. F. (2021). Individual Participant Data Meta-Analysis for Healthcare Research. *Individual Participant Data Meta-Analysis: A Handbook for Healthcare Research,* 1-6.

Rutjes, A. W.S., Reitsma, J. B., Nisio, M. D., Smidt, N., van Rijn, J. C., & Bossuyt, P. M. M. (2006). Evidence of bias and variation in diagnostic accuracy studies. *Canadian Medical Association Journal, 174*(4), 469-476. https://doi.org/ 10.1503/cmaj.050090

Salaschek, M., Zeuch, N., & Souvignier, E. (2014). Mathematics growth trajectories in first grade: Cumulative vs. compensatory patterns and the role of number sense. *Learning and Individual Differences, 35,* 103–112. https://doi.org/10.1016/j.lindif.2014.06.009

Samuels, C. A. (2011). RTI: An approach on the march. *Education Week, 30*(22), 2-5.

Scammacca, N., Fall, A. M., Capin, P., Roberts, G., & Swanson, E. (2020). Examining factors affecting reading and math growth and achievement gaps in grades 1–5: A cohort-sequential longitudinal approach. *Journal of educational psychology*, *112*(4), 718.

Scarborough, H. S. (1998). Predicting the future achievement of second graders with reading disabilities: Contributions of phonemic awareness, verbal memory, rapid naming, and IQ. *Annals of Dyslexia*, *48,* 115-136.

Smolkowski, K. & Cummings, K. D. (2015). Evaluation of diagnostic systems: The selection of students at risk of academic difficulties. *Assessment for Effective Intervention, 41*(1), 41-54. https://doi.org/10.1177/1534508415590386.

Stijnen, T., Hamza, T. H., & Özdemir, P. (2010). Random effects meta‐analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Statistics in medicine*, *29*(29), 3046-3067.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285–1293. https://doi.org/10.1126/science.3287615

Thomas, A. S., & January, S. A. A. (2019). Evaluating the criterion validity and classification accuracy of universal screening measures in reading. *Assessment for Effective Intervention*. https://doi.org/10.1177/1534508419857232

Trevethan, R. (2017). Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. *Frontiers in Public Health, 5*(November), 1–7. https://doi.org/10.3389/fpubh.2017.00307

VanDerHeyden, A. M. (2013). Universal screening may not be for everyone: Using a threshold model as a smarter way to determine risk. *School Psychology Review, 42*(4), 402–414. https://doi.org/10.1080/02796015.2013.12087462

Whiting, P., Rutjes, A. W., Reitsma, J. B., Glas, A. S., Bossuyt, P. M., & Kleijnen, J. (2004). Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Annals of internal medicine*, *140*(3), 189-202.

Yeo, S. (2010). Predicting performance on state achievement tests using curriculum- based measurement in reading: A multilevel meta-analysis. *Remedial and Special Education, 31*(6), 412-422. https://doi.org/10.1177/0741932508327463

Youman, M., & Mather, N. (2018). Dyslexia laws in the USA: A 2018 update. *Perspectives on Language and Literacy*, *44*(2), 37–41. Retrieved from http://www.DyslexiaIDA.org