Journal of School Psychology 108 (2025) 101403

Contents lists available at ScienceDirect

Journal of School Psychology

journal homepage: www.elsevier.com/locate/jschpsyc

# Accounting for learning environments in academic screening

Adam J. Lekwa [a,*], Linda A. Reddy [a], Ryan J. Kettler [a], Ethan R. Van Norman [b]

[a] Department of School Psychology, Rutgers University, United States
[b] Department of Education and Human Services, Lehigh University, United States

ARTICLE INFO

Action Editor: Katie Maki
Editor: Craig A. Albers

*Keywords:*
Screening
Student achievement
Assessment of teaching
Student risk

ABSTRACT

Within multi-tiered systems of support (MTSS) practice and research, students' need for academic intervention is often determined by comparison of students' screening scores to cut scores. We examined the degree to which the relationship between students' fall screening data (i.e., Measures of Academic Progress) and their outcome on a spring summative state test related to the quality of the classroom learning environment and how core instructional strategies influenced this relationship. Fall screening data and spring state test outcomes in English/language arts (ELA) and math were analyzed from a sample of 72 teachers and 1554 third-grade students. Multilevel logistic regression revealed that the association between students' ELA or math skills at the beginning of a school year and state test at the end of the year were not identical across classrooms (odds ratios range = 0.81–0.92). A significant interaction was observed between students' fall ELA screening scores and teachers' instructional strategy use in predicting state test outcomes ($p = .03$). Teacher strategy use was found to be a significant contributor to false positives in screening decisions based on optimal cut scores for ELA ($p = .003$), but not math.

## 1. Introduction

A core feature of multi-tiered systems of support (MTSS) is universal screening in which data from brief assessments are collected from all students to evaluate and improve core instruction and to aid in intervention planning. These data serve not only as estimates of students' current skill levels, but also as indications of their risk of an adverse future educational outcome (e.g., Glover & Albers, 2007; Kettler et al., 2014; Ysseldyke et al., 2023). When students are at risk of poor academic outcomes, educators may decide to deliver targeted (i.e., tier 2) or intensive (i.e., tier 3) intervention, or if large proportions of students are identified as at risk (i.e., class-wide need), to make changes to core instruction (i.e., tier 1; Burns et al., 2008; Fuchs et al., 2017; Kupzyk et al., 2012).

### 1.1. Validating typical interpretations and uses of screening data

Interpretations and uses of students' screening scores are inferences based on incomplete data (Christ & Nelson, 2014). In keeping with the argument-based validity framework, the quality of these inferences depends on the degree to which they can be supported with specific evidence as required by interpretation/use arguments (IUA; Kane, 2013). Especially important in the context of screening assessment are inferences about the meaning of a test score for a student's likely future performance on a test of a similar or related

* Corresponding author at: School of Applied and Professional Psychology, Rutgers University, 152 Frelinghuysen Road, Piscataway, NJ 08854, United States.
*E-mail address:* al928@gsapp.rutgers.edu (A.J. Lekwa).

https://doi.org/10.1016/j.jsp.2024.101403
Received 21 August 2023; Received in revised form 14 November 2024; Accepted 21 November 2024
Available online 29 November 2024

construct (i.e., an extrapolation inference) and about students' need for instruction that is somehow qualitatively different from, and often in addition to, tier 1 core instruction (i.e., need for intervention, which is a decision inference). Measures used for screening are expected to demonstrate sufficiently strong correlations with important criteria such as summative test scores (Silberglitt & Hintze, 2005). Subsequently, when a student receives a low screening score relative to a cut score, educators may infer that the student is unlikely to make sufficient progress to meet achievement standards by the end of the school year without intervention (e.g., Glover & Albers, 2007; Kettler et al., 2014).

Evidence supporting these two inferences often comes in the form of (a) statistical models that generate predictions on criterion measures and (b) analyses of the accuracy of the decisions we make about students' need for intervention. Dozens of studies on a wide variety of measures across grade levels have been conducted to document both forms of evidence. The range of topics is diverse, including predictive relationships and accuracy indices for important outcomes (e.g., January & Klingbeil, 2020; Klingbeil et al., 2015), exploring different approaches to deriving cut-scores (e.g., Ford et al., 2017; Johnson et al., 2009; Klingbeil et al., 2019; Patton et al., 2014), combining scores from multiple screeners to derive estimates of risk (e.g. Van Norman, Nelson, & Klingbeil, 2017; VanMeveren et al., 2020), or use of advanced statistical techniques to classify student risk status (e.g., King et al., 2016; Petscher & Koon, 2020; Van Norman, Klingbeil, & Nelson, 2017).

### 1.2. A gap in screening research

Despite these advancements, key classification accuracy indices reported in research often fall short of ideal suggested criteria—such as criteria that apply to estimates of sensitivity (e.g., $\leq 0.90$; Jenkins et al., 2007), specificity (e.g., $\leq 0.80$; Compton et al., 2010), or Area Under the Curve (AUC; e.g., $\leq 0.80$; Hosmer Jr et al., 2013). In addition, inconsistencies in screening decision accuracy across sites using common measures have been documented (Ford et al., 2017; Grapin et al., 2017; Johnson et al., 2009; Klingbeil et al., 2019; Patton et al., 2014; Thomas & January, 2021). This trend may be partially explained by varying base rates of the targeted trait (i.e., risk) in the population being screened or the cut-scores used for screening decisions (Edwards et al., 2022). Differences in either of these will result in differences in accuracy indices from site to site or across empirical studies using the same screening measure. Yet, evidence in the form of predictive relationships and classification accuracy indices might provide incomplete support for inferences in an IUA for screening: It is possible that some of the differences in screening quality can be accounted for by measures of key aspects of the learning environment at school, such as the quality of teachers' instructional strategy use. Scholars sometimes recommend the use of information about local conditions in estimation of students' risk of adverse outcomes (e.g., Clemens et al., 2016; Klingbeil et al., 2019; VanDerHeyden, 2013). However, to date, studies demonstrating the value of such approaches have only included information about students' prior levels of achievement. Whether, and to what extent, the validity of screening decisions might depend on aspects of instruction has yet to be studied.

#### 1.2.1. Instructional strategy use and screening decisions

Learning environments include factors that may influence student learning and success such as curriculum and the strategies teachers use in delivery of instruction (Carroll, 1963; Nye et al., 2004; Rockoff, 2004; Wang et al., 1993). Specifically, quality of instructional strategy use has been linked to student outcomes (Doabler et al., 2021; Grossman et al., 2014; McLean et al., 2016) as well as differing rates at which students gain achievement in mathematics and reading (Lekwa et al., 2019).

Instructional strategy use is multidimensional; what constitutes effective teaching can vary depending on student skill levels, with students exhibiting low skill levels relative to grade expectations experiencing greater benefits from teachers' use of explicit instructional strategies (Archer & Hughes, 2010). This trend has been documented for mathematics (Burns et al., 2010; Doabler et al., 2021) and reading instruction alike (Burns et al., 2024; Carlisle et al., 2011; Connor, Morrison, & Katch, 2004; Szadokierski et al., 2017). In other words, students at different skill levels benefit in different ways from teachers' instructional practices. Moreover, instructional strategy use has been shown to vary substantially between teachers in a school building and between schools as well (Reddy et al., 2013a; Ritzema et al., 2016). If students experience different gains in response to teachers' instructional strategy use, and if teachers vary in their use of instructional strategies, then actual (versus estimated) levels of students' risk could be associated with how teachers teach, and accuracy of risk classification determined by cut score might differ on this basis. Whether such variation might influence the validity of screening and decisions based on screening scores has yet to be explored empirically.

### 1.3. The present study

Clemens et al. (2016) speculated that a proportion of false positives (i.e., students identified as needing intervention who in truth did not need intervention) might be students who lacked the requisite skills to successfully complete the screener initially but went on to benefit from effective core instruction. Instead of treating students that presumably benefited from tier 1 instruction as inaccurate estimates of risk, Clemens and colleagues recommended that researchers studying academic screening in MTSS attempt to account for instructional effects in the calculation of cut scores or estimation of subsequent classification accuracy. The purpose of the present study was to address this gap with an exploration of whether information about instructional practices, in addition to students' skill levels, could explain variation in outcomes of end of year state tests in reading and mathematics (i.e., extrapolation inferences). We also examined how accuracy of decisions about students' risk—or need for academic intervention—might be related to quality of core (tier 1) instruction that occurs after screening (i.e., decision inferences). Three research questions were examined:

1. Does the predictive relationship between student fall screening scores and probability of proficiency on end-of-year state tests vary between classrooms?
2. Can variability in predictive relationships between student fall screening scores and end-of-year state test scores be explained by the teachers' instructional practices?
3. Is accuracy of screening decisions (intervention needed vs. not needed) related to instructional practices that occur after screening?

## 2. Method

### 2.1. Sample

Data for this study were obtained from three successive cohorts of third-grade students ($N = 1554$) and their teachers ($N = 72$) in 12 charter schools in the mid-Atlantic US between the 2014–15 and 2016–17 school years. Students were 56% female (the number of students who were non-binary was not available in this sample). By race/ethnicity, 48% were Black, 37% Hispanic, 10% White, and 4% Asian (the remaining 1% were recorded as multiracial, Native American, or unspecified). Teachers were mostly female (81%), were 51% White, 32% Black, 1% Native American, or 8% other; demographic data were not available for the final 8% of teachers in the sample. As displayed in Table 1, 44% of students across cohorts met or exceeded standards in English/language arts (ELA) and math alike. These rates were slightly lower than state-wide results that ranged from 44% to 53% between 2015 and 2017.

The sample was drawn from the U.S. Department of Education funded School System Improvement (SSI) Project, a collaborative effort between charter schools and universities to enhance human capital management systems through rigorous teacher evaluation. All schools met federal poverty status by having at least 50% of their students eligible for either free or reduced-price lunch. These schools were not implementing MTSS but were rather operating using the state's required pre-referral intervention procedures that employed a team-based problem-solving model in which classroom teachers referred students of concern to multi-disciplinary teams who recommend intervention plans. This arrangement is common in schools, and data on problem-solving team practices and outcomes in the field suggest needs for improvement in terms of effectiveness and efficiency (Sims et al., 2024).

As part of their participation in the SSI Project, schools collected screening data in reading and math each fall and spring using the Northwest Evaluation Association Measures of Academic Progress (NWEA MAP) tests. In addition, schools collected data on teacher practices using the Classroom Strategies Assessment System (CSAS; Reddy et al., 2013b) three times for each teacher each school year (see Glover et al., 2016). This study was approved by the institutional review board at Rutgers University.

### 2.2. Measures

#### 2.2.1. Northwest Evaluation Association Measures of Academic Progress

Student achievement in reading and math was screened in the fall of each school year with the MAP Reading and Math assessments

**Table 1**
Descriptive statistics.

| Measure | Cohort | Valid | Missing | *M* | *SD* |
|---|---|---|---|---|---|
| *Mathematics* | | | | | |
| Fall MAP Math RIT | a | 399 | 12 | 188.47 | 11.14 |
| | b | 461 | 31 | 179.19 | 11.92 |
| | c | 504 | 33 | 169.89 | 16.87 |
| | Total | 1364 | 76 | 178.47 | 15.68 |
| PARCC Math (% Met or Exceeded) | a | 411 | 0 | 0.35 | 0.48 |
| | b | 492 | 0 | 0.43 | 0.50 |
| | c | 537 | 0 | 0.53 | 0.50 |
| | Total | 1440 | 0 | 0.444 | 0.497 |
| CSAS IS Discrepancy Scores | a | 18 | 4 | 13.81 | 12.42 |
| | b | 25 | 4 | 18.04 | 11.32 |
| | c | 16 | 4 | 13.27 | 8.36 |
| | Total | 59 | 12 | 15.03 | 10.883 |
| *Reading* | | | | | |
| Fall MAP Reading RIT | a | 380 | 64 | 188.17 | 14.38 |
| | b | 471 | 61 | 178.52 | 14.15 |
| | c | 502 | 71 | 171.31 | 17.36 |
| | Total | 1349 | 91 | 178.55 | 16.88 |
| PARCC ELA (% Met or Exceeded) | a | 444 | 0 | 0.39 | 0.49 |
| | b | 532 | 0 | 0.46 | 0.50 |
| | c | 573 | 0 | 0.45 | 0.50 |
| | Total | 1549 | 0 | 0.44 | 0.50 |
| CSAS IS Discrepancy Scores | a | 20 | 4 | 15.57 | 12.28 |
| | b | 30 | 4 | 17.65 | 10.71 |
| | c | 18 | 3 | 12.48 | 7.84 |
| | Total | 68 | 11 | 15.11 | 10.49 |

*Note*. RIT = Rasch Unit (i.e., the score produced by NWEA MAP).

using RIT scores (NWEA, 2011). The MAP are computer adaptive tests of broad reading and math skills for students in kindergarten through 12th grade. Output provides a variety of scores based on student performance. Typical administration times for the MAP range from 15 to 60 min depending on grade level. NWEA vertically scaled the MAP assessment to allow for comparison of individual students' scores both within school years and between grades. The tests were designed to be administered multiple times across a school year to provide schools with descriptions of students' growth in addition to single point estimates of student achievement. Test administrators must complete multiple day trainings involving NWEA Professional Learning Consultants to become proficient in test administration. The 2011 MAP Technical Report (NWEA, 2011) provides a comprehensive plan for ensuring the integrity of results from the assessment.

NWEA (2011) reported estimates of internal consistency >0.90 for MAP Reading and MAP Math across all grades and marginal reliabilities for MAP ranging from 0.94 to 0.97 across grades and content areas. Marginal reliability, which is an index of internal consistency, combines measurement error estimated at different points on the test into a single index. NWEA (2011) also reported strong evidence of concurrent validity (estimates range = 0.67–0.88) and predictive validity (estimates around 0.70) with state tests of achievement in both subjects. These validity coefficients are estimated and reported periodically based on averages across a wide range of state large scale proficiency tests in ELA and math. For example, the 2019 MAP Technical Report (NWEA, 2019) included concurrent and predictive validity indices for 24 US states.

#### 2.2.2. Classroom Strategy Assessment System

Teachers' instructional practices were assessed using the Classroom Strategies Assessment System–Observer Form (CSAS; Reddy et al., 2013b). The CSAS is an observation-based rating scale designed to gather data about qualities of teachers' strategy use in classroom behavior management and instructional delivery. It is completed by conducting a 30-min observation of teaching during which observers record discrete counts of specific instructional and behavior management strategies (i.e., opportunities to respond, directives, praise, and corrective feedback) and targeted notes on qualities of strategy use in five instructional and four behavior management dimensions. Observers use discrete counts of basic strategies as well as targeted notes on multiple aspects of instruction to complete the instrument's two rating scales immediately after each 30-min observation, including the (a) Instructional Strategies rating scale (IS) and the (b) Behavior Management Strategies rating scale (BMS). Only IS scores were used for the purposes of this study.

Research on the CSAS has demonstrated an empirical factor structure that was consistent with the intended structure, as well as adequate internal consistency (Cronbach's α of 0.91 for IS Rating Scale, and 0.92 for BMS Rating Scale), inter-observer agreement (92% agreement for the IS Rating Scale, and 88% agreement for the BMS Rating Scale), and test-retest reliability indices ($r = 0.86$ for the IS Rating Scale and 0.80 for the BMS Rating Scale; Reddy et al., 2013b, 2015). CSAS scores have been found to predict student academic achievement, with an odds ratio of 0.74 for meeting grade level standards in math and 0.73 for reading. In other words, students in classrooms of teachers with higher CSAS discrepancy scores had lower probabilities of reaching grade level standards in math and reading (Reddy et al., 2013c). CSAS scores have also been found to predict gains on state mandated tests of math and reading (Lekwa et al., 2019).

The IS scale, which was included in the present study, consists of 28 items that produce a total score, two composite scores, and five subscale scores. Observers make two ratings to complete each item. First, observers rate how often (Observed Frequency Rating) teachers use specific instructional strategies on a 7-point scale that uses the following anchors: 1 (*never used*), 3 (*sometimes used*), and 7 (*always used*). Second, observers rate how often the teachers *should have* used each strategy (Recommended Frequency) on the same 7-point scale (1 = *never used*, 3 = *sometimes used,* 7 = *always used*). CSAS training gives observers the basis for Recommended Frequency ratings given observed lesson objectives and student functioning. The absolute difference between these two ratings (i.e., discrepancy score) is calculated for each item (i.e., |recommended frequency - frequency ratings|). Discrepancy scores are interpreted as indications of the extent to which any change in strategy use appeared to be necessary. Larger discrepancy scores indicate greater need for change in strategy use; smaller discrepancy scores indicate lesser need for change. Item discrepancy scores are summed across items to create the corresponding composite and total IS discrepancy scores. Teachers' IS discrepancy scores across cohorts in this sample exhibited strong internal consistency, with omega coefficients between 0.94 and 0.96.

#### 2.2.3. Partnership for Assessment of Readiness for College and Career Assessments

The Partnership for Assessment of Readiness for College and Career Assessments (PARCC; Education Testing Services et al., 2016) tests are computer-based tests of student achievement of Common-Core State Standards in ELA and math. Used as mandatory accountability assessment in numerous US states since 2015, PARCC tests provide annual data on students' academic achievement in Grades 3–11. Scores produced by PARCC tests are classified into one of five categories for students in third grade: (1) did not yet meet, (2) partially met, (3) approached expectations, (4) met expectations, and (4) exceeded expectations. For both subjects, scores of 750 and greater indicated students met or exceeded expectations for grade level standards.

Like any large-scale assessment for state accountability reporting, the PARCC test and administration process are heavily manualized and reviewed for quality and comprehensiveness by the U.S. Department of Education. Test site coordinators are responsible for organizing test administrators and ensuring the tests are administered with fidelity and security. Any test irregularities or security breaches are reported to the test coordinator by standardized form within two school days of their occurrence. Reliability estimates for PARCC scores range from $\alpha = 0.85$ to $\alpha = 0.94$ for the math tests and from $\alpha = 0.89$ to $\alpha = 0.93$ for the ELA tests (Education Testing Services et al., 2016). Additional correlational evidence supports the PARCC tests' intended factor structures (Pearson, 2019) and indicates that students' levels of performance on PARCC are highly consistent with their performance on MAP (with consistency rates of 0.84 and 0.85 in ELA/reading and math, respectively; NWEA, 2016).

### 2.3. Procedures

#### 2.3.1. Screening and accountability assessment

Students completed MAP reading and math assessments at the beginning (September) and end of each school year (May) of the SSI Project. Students completed PARCC testing in ELA and math between April and May each project year (2015, 2016, and 2017).

#### 2.3.2. CSAS observer training and data collection

Administrators of participating schools were required to conduct a set of at least three classroom observations using the CSAS for each teacher as part of the formative teacher evaluation process implemented in the SSI Project. Prior to conducting these observations, administrators participated in a 3-day CSAS training; all administrators completed each day of training. This training started with an orientation to main findings from the scientific literature on which the CSAS was based (i.e., research on effective instructional practices; Brophy & Good, 1986; Hattie, 2009; Rosenshine, 2012). Following this orientation, administrators were taught how to complete *observed* and *recommended* frequencies for a set of video-recorded lessons based on (a) the effective instruction literature and (b) targeted notes taken while observing recorded instruction. Training was conducted in phases; each administrator was required to complete knowledge tests at multiple points during training prior to moving on to the next phase of training. After completing the didactic portion of training, administrators practiced coding classroom videos using the CSAS and practice feedback was provided by a CSAS Trainer/Master Coder. Finally, administrators were required to pass a test administration of the CSAS in which they were asked to independently rate five videos of classroom instruction using the CSAS. To pass this test, administrators were required to provide ratings of "observed" and "recommended" frequencies of strategy use that were within one point of those provided by CSAS Trainer/Master Coders on a minimum of 32 out of 54 items. Across administrators, this criterion for performance corresponded with an average intraclass correlation (ICC) (3,*k*) of 0.69 between administrators and Trainers/Master Coders (Bryer, 2023). Qualitatively, this would be considered a "good" (Cicchetti, 1994) or "moderate" (Koo & Li, 2016) level of consistency in ratings. All administrators ($N = 12$) in the present study passed reliability testing and received subsequent co-observation practice in actual classrooms with a certified CSAS Trainer/Master Coder.

Although administrators were expected to complete a minimum of three observations per school year using the CSAS for each teacher in their buildings, there was an average of 2.2 observations conducted per teacher ($SD = 0.88$). A total of 44% of observations were conducted in ELA arts classes, 30% in math classes, and 26% in other classes (i.e., content areas such as science or history). To be included in the dataset, teachers needed to be rostered as instructors of record for students' general education ELA and math classes. These observations occurred throughout each school year (during the fall, winter, and spring) in accordance with the teacher evaluation system procedures used throughout the SSI Project, aligned with state teacher evaluation policy. For the purposes of this study, the total IS discrepancy scores for the three CSAS observations were averaged together for each teacher, respectively, each year (thus, each teacher would have up to three CSAS IS discrepancy scores in the dataset). A teacher's average IS discrepancy score may be interpreted as an estimate of the average quality of instructional strategy use throughout a school year; descriptive statistics for all measures are displayed in Table 1.

### 2.4. Data analyses

#### 2.4.1. Missing data

Fall MAP Reading or Math scores were missing for 13% and 5% of students, respectively. Likewise, teachers' IS discrepancy scores were missing in 14% and 17% of cases in the reading and math datasets, respectively (calculated at the teacher level), or 7% when calculated at the student level for both subjects. No state test scores were missing in this dataset. Analysis of patterns of missingness suggested data were missing at random (MAR) versus missing completely at random (MCAR; Little's MCAR $p < .05$), and multilevel multiple imputation was applied using chained equations with the "mice" package in the R statistical platform (Van Buuren & Groothuis-Oudshoorn, 2011). A set of 100 parallel datasets were generated for both subjects separately. Results of multilevel models fit to these datasets were pooled using Rubin's (2018) rules. Model comparisons involved the generation and pooling of 100 likelihood ratio tests following procedures described by Chan and Meng (2022), which yields an *F* statistic and associated degrees of freedom to evaluate the statistical significance of the superiority of a full versus a reduced (null) model (see Grund et al., 2023, for a review). Necessary assumptions about random effect and residual distributions were confirmed on original, non-imputed data using visual analysis with the R package "DHARMa" (Hartig, 2017).

#### 2.4.2. Research question (RQ) 1: influence of classroom on the relationship between MAP and PARCC scores

To address this question, we fit three-level logistic models of PARCC Math and ELA outcomes (scores of 750 or higher indicating students "met" or "exceeded expectations" versus scores below 750, indicating that students did not meet grade level expectations for achievement in ELA or Math) including the associated math or reading fall screening scores (centered around the grand mean [CGM]) as predictors. Student scores were nested within classrooms, which were nested within schools. A set of uncorrelated random intercepts and MAP slopes were estimated at Level 2 for classrooms and random intercepts were estimated at Level 3 for schools. The magnitude of the random effect for MAP slope captures the heterogeneity in the correlation between initial skill level and end of year proficiency across classrooms in the sample. The significance of the random slope variance estimates was evaluated with pooled likelihood ratio tests (Chan & Meng, 2022).

#### 2.4.3. RQ 2: relationship between MAP and PARCC at different levels of instructional quality

Two additional models were estimated for RQ2. First, teachers' IS discrepancy scores (CGM) as a classroom-level fixed effect was added to the 3-level model proposed for RQ 1. A model that included an interaction term between IS discrepancy scores and students' MAP scores (CWC) was estimated. Each classroom's average MAP score (centered around the grand mean of classroom average MAP scores) was included as an additional Level-2 covariate to aid interpretation of the cross-level interaction between teachers' IS discrepancy scores and students' fall MAP scores. Interested readers may refer to Enders and Tofighi (2007) for rationales for mean-centering and accounting for cluster averages in estimation of fixed effects.

The interaction terms in the final model capture the extent to which the relationship between fall MAP scores and PARCC end-of-year outcomes depended on instructional quality. As with RQ 1, relative model fits were evaluated with pooled results of likelihood ratio tests (Chan & Meng, 2022). The full model, including all main and interaction effects for both subjects is in Eq. (1), below.

$$PARCC_{ijk} = \delta_{000} + \gamma_{01}\overline{MAP}_{jk} + \gamma_{02}IS_{jk} + \gamma_{03}\left(\overline{MAP}_{jk}{*}IS_{jk}\right) + \gamma_{10}MAP_{ijk} + \gamma_{11}\left(MAP_{ijk}{*}IS_{jk}\right) + u_{0jk} + u_{1jk}MAP_{ijk} + u_{00k} + \varepsilon_{ijk} \quad (1)$$

This model included a main effect for fall MAP score (Reading or Math; $\gamma_{10}MAP_{ijk}$) for students ($i$) clustered within classrooms ($j$) and schools ($k$), as well as the effect of teachers' IS discrepancy scores at the classroom level ($\gamma_{02}$), classroom average MAP scores ($\gamma_{01}$) and the cross-level interactions between students' fall MAP scores and teachers' yearly IS scores ($\gamma_{11}$), and between teachers' IS scores and classroom average MAP scores ($\gamma_{01}$). The model's stochastic portion included random intercepts for classroom ($u_{0jk}$), random slopes for students' MAP scores at the classroom level ($u_{1jk}$), and random intercepts for schools ($u_{00k}$), and residuals ($\varepsilon_{ijk}$). RQ1 and RQ2 models were fit in R using the glmmTMB package (Magnusson et al., 2017).

#### 2.4.4. RQ 3: instructional conditions and variation in classification accuracy

To assess the extent to which teachers' instructional practices were associated with incorrect predictions of students' risk status (or need for intervention), we generated cut scores for MAP Reading and Math in prediction of their associated PARCC to identify students as "at risk" or "not at risk." We followed an optimization procedure to identify a cut score that would prioritize sensitivity over specificity. Cut scores were calculated in a two-step process. First, we split students randomly into two datasets. We used the first dataset to identify an optimal cut-score using bootstrapped (1000 samples each subject) Receiver Operating Characteristic (ROC) curve analysis, optimizing sensitivity (Thiele & Hirschfeld, 2020). Next, we cross validated these cut scores on the second dataset. If students' fall MAP Math or fall MAP Reading scores were above their respective cut-scores, we classified those students as "not at risk" in that subject, or as likely to meet grade level expectations without intervention. If MAP Math or Reading scores were below their respective cuts in either subject, we classified students as "at risk". We then compared these classifications with students' actual PARCC outcomes—whether they met or failed to meet grade level expectations. This yielded four possible classifications for each student: True Positive, True Negative, False Positive, and False Negative. Following the same imputation and pooling procedures specified above, we fit a set of multilevel logistic models to the multiply imputed datasets to examine the degree to which teachers' IS discrepancy scores predicted false positive and false negative classification errors. The initial model included linear and quadratic terms for MAP Math and MAP Reading, respectively. Students toward either extreme of the RIT scale are likely to be classified correctly in screening. The initial model was used as a baseline against a second model in which we examined the additive value of teachers' IS discrepancy scores in prediction of false positives and false negatives for each subject. Model fit was compared in the same manner as previous analyses.

## 3. Results

Descriptive statistics for all students' NWEA MAP screening scores, PARCC outcomes, and teachers' CSAS IS discrepancy scores are displayed in Table 1 for each student cohort separately and in total. All results in Tables 2–5 are presented in the log scale; odds ratios reported below were obtained through exponentiation of these values.

**Table 2**
Testing variance of relationship between MAP and PARCC.

| | PARCC Math, A | | PARCC Math, B | | PARCC ELA, A | | PARCC ELA, B | |
|---|---|---|---|---|---|---|---|---|
| Parameter Estimates | *b* (*SE*) | *p* | *b* (*SE*) | *p* | *b* (*SE*) | *p* | *b* (*SE*) | *p* |
| Fixed Effects | | | | | | | | |
| (Intercept) | 0.42 (0.32) | 0.19 | 0.53 (0.36) | 0.14 | 0.63 (0.40) | 0.11 | 0.67 (0.40) | 0.10 |
| Fall MAP Math (CWC) | −0.11 (0.01) | < 0.001 | −0.15 (0.02) | < 0.001 | | | | |
| Fall MAP Math Classroom Mean (CGM) | 0.01 (0.02) | 0.60 | 0.01 (0.02) | 0.46 | | | | |
| Fall MAP Reading (CWC) | | | | | −0.11 (0.01) | < 0.001 | −0.12 (0.01) | < 0.001 |
| Fall MAP Reading Classroom Mean (CGM) | | | | | −0.02 (0.01) | 0.19 | −0.02 (0.01) | 0.15 |
| Random Effects | | | | | | | | |
| Intercept Variance | 0.61 (Teacher) | | 0.68 (Teacher) | | 0.17 (Teacher) | | 0.17 (Teacher) | |
| | 1.01 (School) | | 1.24 (School) | | 1.73 (School) | | 1.78 (School) | |
| Slope Variance (Fall MAP CWC) | | | 4.65E-03 | | | | 6.53E-04 | |

*Note*. CWC = Centered within Cluster; CGM = Centered on Grand Mean.

**Table 3**
Relationship between MAP and PARCC math scores at different levels of instructional quality.

| | PARCC Math, C | | PARCC Math, D | |
|---|---|---|---|---|
| Parameter Estimates | *b* (*SE*) | *p* | *b* (*SE*) | *p* |
| Fixed Effects | | | | |
| (Intercept) | 0.52 (0.37) | 0.15 | 0.55 (0.36) | 0.13 |
| Fall MAP Math (CWC) | −0.15 (0.02) | < 0.001 | −0.15 (0.01) | < 0.001 |
| Class Avg. MAP Math | 0.01 (0.02) | 0.51 | 0.01 (0.02) | 0.63 |
| IS Discrepancy (CGM) | 0.01 (0.02) | 0.49 | 0.02 (0.02) | 0.30 |
| MAP Math (CWC) * IS | | | 8.16E-04, (1.09E-03) | 0.45 |
| Avg. MAP Math * IS | | | −1.43E-03, (1.92E-03) | 0.46 |
| Random Effects | | | | |
| Intercept Variance | 0.67 (Teacher) | | 0.68 (Teacher) | |
| | 1.30 (School) | | 1.27 (School) | |
| Slope Variance (Fall MAP CWC) | 4.51E-03 | | 4.42E-03 | |

*Note*. CWC = Centered within Cluster; CGM = Centered on Grand Mean.

**Table 4**
Relationship between MAP and PARCC reading scores at different levels of instructional quality.

| | PARCC ELA, C | | PARCC ELA, D | |
|---|---|---|---|---|
| Parameter Estimates | *b* (*SE*) | *p* | *b* (*SE*) | *p* |
| Fixed Effects | | | | |
| (Intercept) | 0.67 (0.43) | 0.12 | 0.69 (0.44) | 0.12 |
| Fall MAP Reading (CWC) | −0.12 (0.01) | < 0.001 | −0.12 (0.01) | < 0.001 |
| Class Avg. MAP Reading | −0.03 (0.01) | 0.08 | −0.03 (0.02) | 0.06 |
| IS Discrepancy (CGM) | 0.03 (0.01) | 0.01 | 0.04 (0.01) | 0.003 |
| MAP Rdg (CWC) * IS | | | −1.91E-03 (−7.64E-04) | 0.01 |
| Avg. MAP Rdg * IS | | | −1.06E-03 (1.28E-03) | 0.41 |
| Random Effects | | | | |
| Intercept Variance | 0.10 (Teacher) | | 0.11 (Teacher) | |
| | 2.06 (School) | | 2.15 (School) | |
| Slope Variance (Fall MAP CWC) | 6.78E-04 | | 4.98E-04 | |

*Note*. CWC = Centered within Cluster; CGM = Centered on Grand Mean.

### *3.1. RQ 1: influence of classroom on the relationship between MAP and PARCC scores*

Results of multilevel logistic models of PARCC Math and ELA outcomes (students meeting or exceeding achievement expectations versus those who did not meet, partially met, or approached expectations) can be found in Table 2. Two models were fit for each subject: Model A, containing fall MAP scores as a fixed effect, and random intercepts at the classroom level, and Model B, which was the same, but also included random slopes for MAP at the classroom level. MAP RIT scores obtained in the fall of each school year were significantly related to spring PARCC scores in Math ($F$ (2, 63,189) = 166.89, $p < .001$) and ELA ($F$(2, 12,180) = 199.51, $p < .001$). Random slopes for MAP (Model B) were significant in both subjects (Math: $F$ (1, 37,618) = 41.83, $p < .001$; Reading: $F$ (1, 119,600) = 4.27, $p = .02$), suggesting that the true relationship between MAP and PARCC varied across classrooms. For PARCC Math, the odds that students would fail to meet expectations decreased by a factor of 0.83 at each higher point on the MAP RIT scale. Considering a slope variance estimate of approximately 0.005, the odds ratio between fall MAP Math and PARCC varied from 0.81 to 0.92 (+/− 1 *SD*). The average odds ratio between fall MAP Reading and PARCC ELA was estimated to be 0.89, which also varied slightly across classrooms with odds ratios ranging from 0.87 to 0.91 (+/− 1 *SD)*.

### *3.2. RQ 2: relationship between MAP and PARCC at different levels of instructional quality*

Teachers' CSAS IS discrepancy scores were not significantly related to students' PARCC outcomes in Math after controlling for students' fall MAP RIT scores (CWC) and classroom average fall MAP RIT scores (Table 3; $F$(1, 1.59e+05) = 0.48, $p = .49$; Math Model C). Teachers' IS discrepancy scores significantly predicted PARCC ELA outcomes after controlling for student (CWC) and classroom average fall MAP Reading RIT scores (Table 4; $F$(1, 35,266) = 5.88, $p = .02$; ELA Model C). There was a significant negative interaction between IS discrepancy scores and MAP Reading scores ($F$(1, 37,941) = 3.57, $p = .03$; ELA Model D), such that students in classrooms with *higher* IS discrepancy scores (greater need for change in instruction) experienced greater probabilities of obtaining a PARCC score below the "met" or "exceeded expectations" range. The effect of instructional quality was stronger among students with below average fall MAP RIT scores (Fig. 1).

**Table 5**
Probability of classification errors as a function of IS discrepancy score.

| | PARCC Math | | | | PARCC English/Language Arts | | | |
|---|---|---|---|---|---|---|---|---|
| | False Positives | | False Negatives | | False Positives | | False Negatives | |
| Parameter Estimates | *b* (*SE*) | *p* | *b* (*SE*) | *p* | *b* (*SE*) | *p* | *b* (*SE*) | *p* |
| Fixed Effects | | | | | | | | |
| (Intercept) | −1.40 (0.31) | < 0.001 | −6.40 (1.01) | < 0.001 | −1.61 (0.47) | < 0.001 | −4.53 (0.73) | < 0.001 |
| Fall MAP Math (CWC) | 2.00E-03 (0.02) | 0.92 | 0.45 (0.11) | < 0.001 | | | | |
| Fall MAP Math (CWC)$^2$ | −5.66E-03 (7.24E-04) | < 0.001 | −0.02 (4.96E-03) | < 0.001 | | | | |
| MAP Math Class Avg. | −0.11 (0.02) | < 0.001 | 0.44 (0.09) | < 0.001 | | | | |
| Fall MAP Reading (CWC) | | | | | −0.02 (0.02) | 0.35 | 0.39 (0.09) | < 0.001 |
| Fall MAP Reading (CWC)$^2$ | | | | | −7.57E-03 (9.27E-04) | < 0.001 | −0.02 (3.83E-03) | < 0.001 |
| MAP Reading Class Avg. | | | | | −0.09 (0.02) | < 0.001 | 0.30 (0.06) | < 0.001 |
| IS Discrepancy (CGM) | −0.01 (0.02) | 0.6 | −2.28E-04 | 0.99 | −0.04 (0.02) | 0.006 | 0.02 (0.02) | 0.38 |
| Random Effects | | | | | | | | |
| Intercepts | 0.87 (Teacher) | | 1.11 (Teacher) | | 0.04 (Teacher) | | 1.22 (Teacher) | |
| | 0.69 (School) | | 0.41 (School) | | 2.14 (School) | | 0.14 (School) | |
| Slope Variance (Fall MAP CWC) | 0.01 | | 0.13 | | 0.01 | | 0.08 | |

*Note*. CWC = Centered within Cluster; CGM = Centered on Grand Mean.

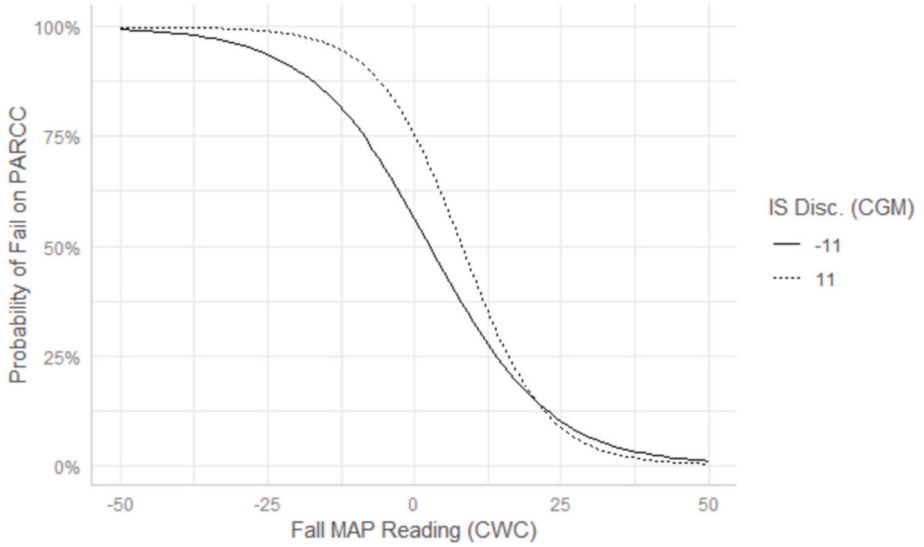


**Fig. 1.** Probability of failing to meet expectations on PARCC ELA by fall MAP reading & IS discrepancy score.

### *3.3. RQ 3: instructional conditions and variation in classification accuracy*

Results of bootstrapped ROC analysis (1000 boot runs) optimizing for sensitivity resulted in a cut score of 186 for the MAP Reading calibration dataset, with an AUC of 0.77, overall classification accuracy of 0.70, sensitivity of 0.83, and specificity of 0.53. For the MAP Math dataset, the optimal cut score (again, prioritizing sensitivity) was 188, with an AUC of 0.75, overall classification accuracy of 0.73, sensitivity of 0.86, and specificity of 0.54.

#### *3.3.1. False positives*

Various students in the cross-validation sample were classified as being "at risk" but who subsequently received PARCC scores above 750 ("met" or "exceeded" expectations) in ELA or math, indicating students met or exceeded grade level expectations (i.e., a false positive error). In the original dataset prior to imputation, this resulted in 89 students in reading and 152 students in math. Pooled results of 3-level logistic regression models included random effects for intercept and slopes for fall MAP RIT at Level 2. As displayed in Table 5, there was a significant quadratic relationship between MAP Math and MAP Reading scores and the corresponding odds of being a false positive (students with extremely low or extremely high fall MAP scores were not likely to be misclassified). Teachers' average IS discrepancy scores were significantly related to students' odds of being false positives on PARCC ELA, $F$ (1, 81,671) = 8.94, $p$ = .003. IS discrepancy scores were not significantly related to false positive status for PARCC Math (Fig. 2).

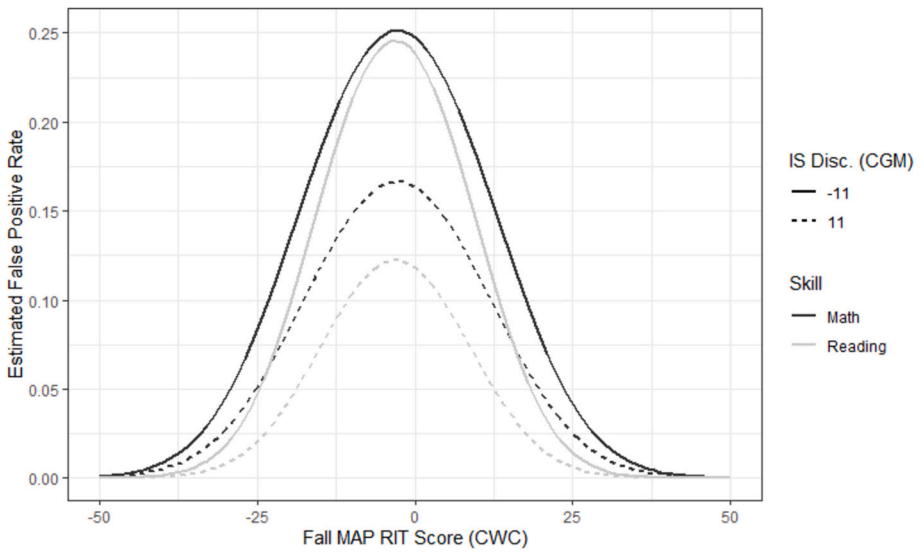


**Fig. 2.** Estimated false positive rate by IS discrepancy score.

### 3.3.2. False negatives

Various students were also identified as false negatives in the sample. In the original data set, prior to imputation, there were 54 students in reading and 46 students in math who were false negatives. Students' fall MAP RIT scores in both subjects were significantly related to the odds of being mis-classified as "not at risk" (Table 5). Teachers' IS discrepancy scores were not significantly related to the odds of being identified as a false negative in either subject.

## 4. Discussion

The purpose of this study was to examine two key inferences that underlie the interpretation and use of screening data, which included extrapolation from a screening score to expected outcomes on a summative assessment, as well as decisions about students' needs for intervention. Although valid interpretation and use of screening data require these inferences, the overall question we addressed in this study was whether these inferences are equally well supported across learning environments that differ in terms of instructional quality. Based on significant random effects estimates reported for RQ 1, we cannot conclude that the relationship between a student's skill level at the beginning of the school year and whether they meet or exceed grade level expectations near the end of the school year is the same across classrooms. Significant variation was found in this relationship for both subjects (reading and math). The effect in this sample was stronger for math than reading. For example, these data suggested that although a student with a fall MAP Math score of 179 (the sample grand mean) would have about a 63% chance of failing to achieve at grade-level expectations on PARCC Math, more realistic estimates for that student would depend on their learning environment (i.e., classrooms and schools) and thus estimated probabilities of failure to meet grade level expectations might vary widely across settings (95% prediction interval of 0.17–0.93).

Moreover, among the variety of factors that might influence student achievement over the course of a school year, we found that teachers' use of instructional practices moderated the relationship between fall screening scores and spring state test outcomes in ELA. The same trend was not apparent for math after controlling for classroom average achievement levels. Additional factors, including curriculum (rigor, scope, and sequence), classroom management practices, or quality of pre-service or in-service training might aid explanation of the varied relationship between fall and spring across classrooms (Andrietti & Su, 2019; McGaffrey et al., 2001; Tarr et al., 2008). Finally, we found that quality of teachers' instructional strategy use, as measured by the CSAS, was a significant predictor of Type 1 error (i.e., false positives) in screening decisions in ELA. These findings are consistent with prior validity studies on the CSAS (Lekwa et al., 2019; Reddy et al., 2013c) in which teachers' CSAS scores were found to be predictive of rates of gain in academic achievement as well as state test outcomes.

### 4.1. Implications for student learning

Clemens et al.'s (2016) summary of research and best practices in screening included discussion of some gaps in screening research, including consideration of environmental variables, and incorporating them into calculation of cut scores for screening decisions. Based on the current set of results, it is possible that inclusion of data like IS discrepancy scores in statistical models would shift estimates of optimal cut scores in small amounts. Regardless, generalizability of cut-scores is limited by several factors such as base rates and correlations between screening instruments and the criterion they predict (Edwards et al., 2022). Upward or downward adjustment of cut-scores alone may be insufficient for promoting more accurate instructional decisions based on screening data. For example, had we applied cut scores published for PARCC by NWEA (which were higher than those we calculated: Third-grade fall RIT scores of 205 in reading and 208 in math; NWEA, 2016), the observed sensitivity indices would have approached 1.0 and specificity indices would have fallen below 0.25; consequently, a majority of students in this dataset would have been flagged as needing intervention. In addition, the cut scores published by NWEA corresponded with the 70th national percentile rank in reading and 65th national percentile in math; cut scores calculated for this study were both at or near the 45th national percentile rank, further illustrating the contextual nature of screening decisions for academic skills. In basing screening decisions primarily on the comparison of students' scores to empirically derived cut-scores, wherever the cut is placed, educators may inadvertently fail to examine the degree to which instruction within any tier does or does not align with students' instructional needs. In other words, screening practices along these lines might not promote effective instruction or intervention as intended. This has implications not only for use of school resources, but also for student learning.

Based on current data, we could estimate that classrooms with the lowest quartile IS discrepancy scores (at or below 11, representing low needs for change in strategy use) would have false-positive screening rates around 25% for ELA when decisions are based primarily on cut-score comparisons whereas students in classrooms with the highest quartile IS discrepancy scores, representing a greater needs for change in instruction, saw ELA false positive rates closer to 10%. In the latter case, lower false positive rates may be interpreted as the result of the use of lower quality instructional strategies after screening and therefore students requiring remediation of prerequisite skills. If we consider a classroom with 20 students, and below-average quality in use of instructional strategies (or the top quartile of CSAS IS discrepancy scores), we might expect about five students to be selected to receive reading intervention that they would not actually need, which is in addition to the number of students identified correctly for such intervention. Spread across multiple classrooms in a grade level, or in a school, the number of students receiving some form of academic intervention unnecessarily could represent a significant loss of opportunity to learn (time spent with instruction not aligned with instructional needs), a burden on school resources (perhaps a greater number of interventionists required), and may negatively impede intervention delivery for students with greater need for intervention. Although this is a scenario that some have suggested is self-correcting assuming use of progress monitoring data (Jenkins et al., 2007), spending time on instruction that is poorly matched with student needs should be

prevented.

Whereas an ecological view of learning is supposed to be a key tenet of MTSS (e.g., Ikeda et al., 2007), it might be weakly incorporated in many schools' implementation of the framework. For example, Briesch et al. (2022) found that approximately 54% of a large sample of school administrators reported determination of students' risk status and need for intervention by comparisons between screening and cut scores (versus intervention assignment decided independently by teachers or teams). In a separate survey, Silva et al. (2021) found that fewer than half of the 387 US-based school psychologists surveyed said their schools used screening data to inform tier 1 instruction. It may be that failure to account for important elements of the learning environment—such as teachers' use of instructional strategies, as in this study—is an unintentional return to thinking about learning problems only as student characteristics. In setting up screening procedures that classify students in terms of risk, educators might inadvertently engage in categorical thinking (i.e., a "tier 2 student" should receive a "tier 2 intervention") and fundamental attribution error (perceiving a person's behavior as determined more by individual characteristics than context; Ross, 1977) instead of skills and opportunities for skills growth. Sabnis et al. (2020) described this potential pitfall in implementation of MTSS in a qualitative analysis of interviews with a group of experienced elementary educators, stating:

> Once placed [in intervention], teachers referred to these students as 'tier 2 students' and 'tier 3 students'. In this way, the tiers came to be a typology of ability in a way that paralleled the traditional typology of special education student and general education student that RTI [MTSS] proponents meant to eliminate. (pg. 296)

Sabnis and colleagues further cautioned about the danger such trends might pose for students in marginalized populations: Lack of attention to instructional variables may prolong or exacerbate longstanding educational inequities, such as disproportionate representation in special education and continued achievement gaps (Farkas et al., 2020).

### 4.2. Screening to inform resource allocation, not student classification

Planning and implementing instruction in any tier of a MTSS framework requires identification of instructional needs more than it requires identification of students. Stating that a student "needs academic intervention" is parallel to stating that there is a mismatch between the instruction soon to be provided (the "what" and "how" of core, or tier 1 instruction) and the student's current skill levels. Bearing in mind that "need for intervention" is a question of alignment between instruction and students' instructional needs, we might better base screening decisions on students' skills relative to grade-level peers and the school's *instructional resources* (i.e., proportion of students that could feasibly receive targeted or intensive intervention; see Kilgus & Eklund, 2016). Schools conduct screening to survey the breadth of instructional needs presented by students, but whether a student needs academic intervention is not a question of their current skill level as much as it is a question of whether core instruction will or will not align with their instructional needs. A majority of students might receive screening scores below an optimal cut point (as was the case in this study), in which case a school should make appropriate changes within core instruction rather than try to implement targeted or intensive interventions for such a large number of students (Parisi et al., 2014).

Similarly, information about teachers' use of instructional strategies may also be helpful for instructional planning and supporting teachers. Data from the present study suggested that core instruction would be effective for a portion of students whose fall screening scores appeared below an optimal cut-score in math or ELA and that an important supportive factor in these cases may be teachers' use of instructional strategies. We found evidence of that relationship for ELA in which some students with equal reading skills at the beginning of a school year experienced substantially different outcomes by the end of the year, and this difference was associated with observers' ratings of teachers' need for change in strategy use. This result is consistent with findings that qualities of teachers' instructional strategy use are associated with students' rates of gain in achievement (Lekwa et al., 2019). Students in classrooms with stronger instruction may experience greater gains and would be more likely to appear as "false negatives" in screening.

Although the IS discrepancy scores do not convey information about the frequency with which specific instructional strategies were used, data from Connor, Morrison, and Katch (2004) supported more specific hypotheses. In a sample of 108 first-grade students and 42 teachers, Connor and colleagues observed interactions between teachers' use of different instructional strategies and students' decoding and vocabulary skills at the beginning of a school year. Students with initially lower levels of skill demonstrated greater gains in classrooms with greater use of explicit (versus implicit) and teacher-managed (versus child-managed) activities, whereas students with greater beginning proficiency experienced greater gains in classrooms with more implicit and child-managed learning activities. Such results highlight that quality of teachers' instructional strategy use is context-dependent and data on strategy use have value for instructional planning and supporting teachers with resources such as data-based coaching (Kretlow & Bartholomew, 2010; Reddy et al., 2021).

Intervention resources should be distributed based not solely on student data or instructional variables but based on both sets of indicators and a combination of the two. To ignore these relationships between teachers' instruction and students' instructional needs would be to continue to situate low achievement within the student and to proceed as if all instructional practices work equally well; both the current findings and educational theory indicate differently.

### 4.3. Limitations

Findings from this study should be interpreted within the context of its limitations. First, participant characteristics may limit generalizability of findings to other teachers, student populations, school contexts, regions, and states. Replication based on diverse samples, schools, and regions would further test and clarify findings presented here. Second, we used an assessment that measured

evidence-based instructional and behavioral management practices that consisted of a range of strategies known to be related to student outcomes. However, we examined a CSAS composite score (i.e., IS discrepancy) and did not assess the associations of specific strategies (e.g., direct instruction, metacognitive strategies) with students' skills and achievement. These relationships also warrant future investigation.

The data analyzed in this study were the products of a school implementation project with school personnel in the context of practice, rather than prospective research. As a result, there are issues that detract from the number and strength of conclusions that can be drawn from them. The IS discrepancy scores used in this study were the average of multiple CSAS administrations per teacher per year (these observations were conducted as part of schools' teacher evaluation systems). These averages do not convey the extent to which teaching could vary over time or across students. Although teachers' annual average CSAS discrepancy scores have been associated with student achievement in this study (for ELA only) and prior studies (Reddy et al., 2013c), it is likely that more conclusive information about the relationship between teacher strategy use and student learning could be obtained from designs in which measures of teaching and learning would be taken in much closer proximity and therefore avoiding the loss of information incurred by averaging over subject areas and an entire school year.

Although administrators were required to successfully complete a training and certification process to administer the CSAS, and internal consistency estimates for this sample of CSAS data were favorable, inter-rater reliability data were unavailable and the degree to which inconsistencies in administrators' use of the CSAS could have impacted results observed here is unknown. More conclusive information about the relationships between variation in teaching strategy use and validity of instructional decision making could be obtained from more tightly controlled replication of this study, including more stringent checks on procedural integrity (such as aligning subject area of classroom observation with subject areas tested), and documentation of inter-rater reliability. Finally, although students' MAP and PARCC scores in reading and math were linked with the teacher responsible for instruction in either area, and schools within this sample were not implementing MTSS frameworks, additional data on any intervention or remediation efforts was not available for analysis. Such effects, if present, might in part explain random intercept and slope variance seen in model results.

### *4.4. Future research*

The present investigation was the first to examine the relationship between a students' skill level at the beginning of the year and proficiency on a summative assessment at the end of the year across classrooms and to test whether differences in learning environments influence this relationship. Specifically, we tested the association of variability in quality of teacher strategy use on the accuracy of prediction of proficiency as indicated by state large scale assessment. In this initial study, quality of teacher strategy use was defined broadly as the discrepancy between ideal and observed use of empirically supported instructional strategies. Thus, further investigations are needed that assess the influence of instructional conditions (e.g., opportunities to learn) and classroom- or student-related factors (e.g., positive classroom management, level of disruption, student academic engagement) on the relationship between screened skill levels and performance. Investigation of these relationships should also examine the impact of lesson formats on quality of core instructional delivery and overall learning environment. Likewise, additional validation work is warranted that measures quality of core instructional practices between reading and math lessons. Future research might also address longer term longitudinal relationships with more data reflecting the trajectory of performance in reading and math. Although we addressed prediction of performance across an academic year, the models did not incorporate previous screening and achievement data, nor did they project into the future beyond an academic year. Screening data should regularly be contextualized within the incremental validity added to data already available to the school system from previous years. The use of statewide achievement proficiency at the end of the academic year as a gold standard implies a certain finality to that point in time, even though performance in reading and mathematics could be very different if measured 3 months later during the summer or 6 months later within the influence of a new learning environment. More longitudinal designs can address the reality that these achievement levels are fluid.

Finally, for schools and educators to adopt the perspective that students' need for intervention is truly a function of the alignment between the content and delivery of instruction and students' instructional needs, additional research and development work should be conducted to enable objective assessment and quantification of instructional match. Such work might open new avenues in screening practices, data interpretation, and preparation and professional development for educators.

## 5. Conclusion

The present investigation empirically tested a long-standing assumption that the relationship between student risk status as determined by academic screening in the fall and proficiency as determined by state standards-based assessments in the spring are similar across classroom environments. This assumption, implicit in conventional use of screening data, ignores another widely held truth that teaching matters. Based on these results, we advocate that decisions about students' needs for intervention should consider the contribution of classroom environment—such as quality of instructional delivery—on the relationship between student skill levels and their ability to meet grade level standards. Doing so recognizes that the screening of risk and need for intervention focuses on neither the student nor the learning environment alone; it is most appropriately focused on the relationship between the two.

## Author note

The current study was implemented as part of the School System Improvement (SSI) Project, a collaboration between multiple universities and charter schools funded by the U.S. Department of Education's Office of Innovation and Improvement as part of the

Teacher Incentive Fund program (awarded to Rutgers, The State University of New Jersey; #S374A120060). The positions and opinions expressed in this article are solely those of the author.

## CRediT authorship contribution statement

**Adam J. Lekwa:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Linda A. Reddy:** Writing – review & editing, Writing – original draft, Funding acquisition. **Ryan J. Kettler:** Writing – review & editing, Writing – original draft, Funding acquisition. **Ethan R. Van Norman:** Writing – review & editing, Writing – original draft, Formal analysis, Conceptualization.

## References


Andrietti, V., & Su, X. (2019). Education curriculum and student achievement: Theory and evidence. *Education Economics, 27*(1), 4–19.

Archer, A. L., & Hughes, C. A. (2010). *Explicit instruction: Effective and efficient teaching*. Guilford Publications.

Briesch, A. M., Chafouleas, S. M., Dineen, J. N., McCoach, D. B., & Donaldson, A. (2022). School building administrator reports of screening practices across academic, behavioral, and health domains. *Journal of Positive Behavior Interventions, 24*(4), 266–277.

Brophy, J. E., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research in teaching* (3rd ed., pp. 328–375). MacMillan.

Bryer, J. (2023). IRRsim: simulate inter-rater reliability statistics. http://irrsim.bryer.org. http://github.com/jbryer/IRRsim. https://jbryer.github.io/IRRsim/.

Burns, M., Duesenberg-Marshall, M., Sussman-Dawson, K., Romero, M., Wilson, D., & Felten, M. (2024). Effects of targeting reading interventions: Testing a skill-by-treatment interaction in an applied setting. *Preventing School Failure. Alternative Education for Children and Youth, 68*(2), 113–121. https://doi.org/10.1080/1045988X.2023.2177982

Burns, M. K., Codding, R. S., Boice, C. H., & Lukito, G. (2010). Meta-analysis of acquisition and fluency math interventions with instructional and frustration level skills: Evidence for a skill-by-treatment interaction. *School Psychology Review, 39*(1), 69–83.

Burns, M. K., VanDerHeyden, A. M., & Boice, C. H. (2008). Best practices in intensive academic interventions. In A. Thomas, & J. Grimes (Eds.), *Best practices in school psychology* (pp. 1151–1162). National Association of School Psychologists.

Carlisle, J., Kelcey, B., Berebitsky, D., & Phelps, G. (2011). Embracing the complexity of instruction: A study of the effects of teachers' instruction on students' reading comprehension. *Scientific Studies of Reading, 15*(5), 409–439.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*(8), 1–9.

Chan, K. W., & Meng, X.-L. (2022). Multiple improvements of multiple imputation likelihood ratio tests. *Statistica Sinica, 32*, 1489–1514. https://doi.org/10.5705/ss.202019.0314.

Christ, T. J., & Nelson, P. M. (2014). Developing and evaluating screening systems: Practical and psychometric considerations. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings: Evidence-based decision making for schools* (pp. 79–110). American Psychological Association. https://doi.org/10.1037/14316-004.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284.

Clemens, N. H., Keller-Margulis, M. A., Scholten, T., & Yoon, M. (2016). Screening assessment within a multi-tiered system of support: Current practices, advances, and next steps. In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.), *Handbook of response to intervention: The science and practice of multi-tiered systems of support* (2nd ed., pp. 187–213). Springer Science + Business Media https://doi.org/10.1007/978-1-4899-7568-3_12.

Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., Cho, E., & Crouch, R. C. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology, 102*(2), 327–340. https://doi.org/10.1037/a0018448.

Connor, C. M., Morrison, F. J., & Katch, L. E. (2004). Beyond the reading wars: Exploring the effect of child-instruction interactions on growth in early reading. *Scientific Studies of Reading, 8*(4), 305–336. https://doi.org/10.1207/s1532799xssr0804_1.

Doabler, C. T., Clarke, B., Kosty, D., Fien, H., Smolkowski, K., Liu, M., & Baker, S. K. (2021). Measuring the quantity and quality of explicit instructional interactions in an empirically validated tier 2 kindergarten mathematics intervention. *Learning Disability Quarterly, 44*(1), 50–62.

Education Testing Services, Pearson, & Measured Progress. (2016). *PARCC: Final technical report for 2015 administration*. Author.

Edwards, A. A., van Dijk, W., White, C. M., & Schatschneider, C. (2022). Screening screeners: Calculating classification indices using correlations and cut-points. *Annals of Dyslexia, 72*(3), 445–460. https://doi.org/10.1007/s11881-022-00261-5.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121.

Farkas, G., Morgan, P. L., Hillemeier, M. M., Mitchell, C., & Woods, A. D. (2020). District-level achievement gaps explain black and Hispanic overrepresentation in special education. *Exceptional Children, 86*(4), 374–392.

Ford, J. W., Missall, K. N., Hosp, J. L., & Kuhle, J. L. (2017). Examining oral passage reading rate across three curriculum-based measurement tools for predicting grade-level proficiency. *School Psychology Review, 46*(4), 363–378.

Fuchs, L. S., Fuchs, D., & Malone, A. S. (2017). The taxonomy of intervention intensity. *Teaching Exceptional Children, 50*(1), 35–43.

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45*(2), 117–135.

Glover, T. A., Reddy, L. A., Kettler, R. J., Kurz, A., & Lekwa, A. J. (2016). Improving high-stakes decisions via formative assessment, professional development, and comprehensive educator evaluation: The school system improvement project. *Teachers College Record, 118*(14).

Grapin, S. L., Kranzler, J. H., Waldron, N., Joyce-Beaulieu, D., & Algina, J. (2017). Developing local oral reading fluency cut scores for predicting high-stakes test performance. *Psychology in the Schools, 54*(9), 932–946.

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher, 43*(6), 293–303.

Grund, S., Lüdtke, O., & Robitzsch, A. (2023). Pooling methods for likelihood ratio tests in multiply imputed data sets. *Psychological Methods, 28*(5), 1207–1221. https://doi.org/10.1037/met0000556.

Hartig, F. (2017). *DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models*. R package version 0.1 (p. 5).

Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Ikeda, M. J., Rahn-Blakeslee, A., Niebling, B. C., Gustafson, J. K., Allison, R., & Stumme, J. (2007). The heartland area education agency 11 problem-solving approach: An overview and lessons learned. In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.), *Handbook of response to intervention: The science and practice of assessment and intervention* (pp. 255–268). Springer Science + Business Media. https://doi.org/10.1007/978-0-387-49053-3_19.

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., 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.

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.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.
Kettler, R. J., Glover, T. A., Albers, C. A., & Feeney-Kettler, K. A. (2014). *Universal screening in educational settings: Evidence-based decision making for schools*. American Psychological Association.
Kilgus, S., & Eklund, K. (2016). Consideration of base rates within universal screening for behavioral and emotional risk: A novel procedural framework. *School Psychology Forum, 10*, 120–130.
King, K. R., Lembke, E. S., & Reinke, W. M. (2016). Using latent class analysis to identify academic and behavioral risk status in elementary students. *School Psychology Quarterly, 31*(1), 43–57. https://doi.org/10.1037/spq0000111.
Klingbeil, D. A., Maurice, S. A., Van Norman, E. R., Nelson, P. M., Birr, C., Hanrahan, A. R., … Lopez, A. L. (2019). Improving mathematics screening in middle school. *School Psychology Review, 48*(4), 383–398.
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. https://doi.org/10.1002/pits.21839.
Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.
Kretlow, A. G., & Bartholomew, C. C. (2010). Using coaching to improve the fidelity of evidence-based practices: A review of studies. *Teacher Education and Special Education, 33*(4), 279–299.
Kupzyk, S., Daly, E. J., III, Ihlo, T., & Young, N. D. (2012). Modifying instruction within tiers in multitiered intervention programs. *Psychology in the Schools, 49*(3), 219–230.
Lekwa, A. J., Reddy, L. A., Dudek, C. M., & Hua, A. N. (2019). Assessment of teaching to predict gains in student achievement in urban schools. *School Psychology, 34*(3), 271–280. https://doi.org/10.1037/spq0000293.
Magnusson, A., Skaug, H., Nielsen, A., Berg, C., Kristensen, K., Maechler, M., et al. (2017). *Package 'glmmTMB'*. R Package Version 0.2 (p. 0).
McGaffrey, D. F., Hamilton, L. S., Stecher, B. M., Klein, S. P., Bugliari, D., & Robyn, A. (2001). Interactions among instructional practices, curriculum, and student achievement: The case of standards-based high school mathematics. *Journal for Research in Mathematics Education, 32*(5), 493–517.
McLean, L., Sparapani, N., Toste, J. R., & Connor, C. M. (2016). Classroom quality as a predictor of first graders' time in non-instructional activities and literacy achievement. *Journal of School Psychology, 56*, 45–58. https://doi.org/10.1016/j.jsp.2016.03.004.
Northwest Evaluation Association (NWEA). (2011). *Technical manual for measures of academic progress (MAP®) and measures of academic progress for primary grades (MPG®)*. Author.
NWEA. (2016). *Linking the PARCC assessments to NWEA MAP tests*. Author.
NWEA. (2019). *MAP growth technical report*. Author.
Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*, 237–257.
Parisi, D. M., Ihlo, T., & Glover, T. A. (2014). Screening within a multitiered early prevention model: Using assessment to inform instruction and promote students' response to intervention. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings: Evidence-based decision making for schools* (pp. 19–46). American Psychological Association. https://doi.org/10.1037/14316-002.
Patton, K. L. S., Reschly, A. L., & Appleton, J. (2014). Curriculum-based measurement as a predictor of performance on a state assessment: Diagnostic efficiency of local norms. *Educational Assessment, 19*(4), 284–301.
Pearson. (2019). *PARCC: Final technical report for 2018 administration*. Author.
Petscher, Y., & Koon, S. (2020). Moving the needle on evaluating multivariate screening accuracy. *Assessment for Effective Intervention, 45*(2), 83–94. https://doi.org/10.1177/1534508418791740.
Reddy, L. A., Dudek, C. M., Fabiano, G., & Peters, S. (2015). Measuring teacher self-report on classroom practices: Construct validity and reliability of the classroom strategies scale–teacher form. *School Psychology Quarterly, 30*, 513–533.
Reddy, L. A., Fabiano, G., Dudek, C. M., & Hsu, L. (2013b). Development and construct validity of the classroom strategies scale-observer form. *School Psychology Quarterly, 28*(4), 317–341. https://doi.org/10.1037/spq0000041.
Reddy, L. A., Fabiano, G. A., Dudek, C. M., & Hsu, L. (2013a). Predictive validity of the classroom strategies scale-observer form on statewide testing scores: An initial investigation. *School Psychology Quarterly, 28*(4), 301–316. https://doi.org/10.1037/spq0000041.
Reddy, L. A., Fabiano, G. A., Dudek, C. M., & Hsu, L. (2013c). Predictive validity of the classroom strategies scale—Observer form on statewide testing scores: An initial investigation. *School Psychology Quarterly, 28*(4), 301–316. https://psycnet.apa.org/doi/10.1037/spq0000041.
Reddy, L. A., Shernoff, E., & Lekwa, A. (2021). A randomized controlled trial of instructional coaching in high-poverty urban schools: Examining teacher practices and student outcomes. *Journal of School Psychology, 86*, 151–168. https://doi.org/10.1016/j.jsp.2021.04.001.
Ritzema, E. S., Deunk, M. I., & Bosker, R. J. (2016). Differentiation practices in grade 2 and 3: Variations in teacher behavior in mathematics and reading comprehension lessons. *Journal of Classroom Interaction, 51*(2), 50–72.
Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247–252.
Rosenshine, B. (2012). Principles of instruction: Research-based strategies that all teachers should know. *American Educator, 36*(1), 12–39.
Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in Experimental Social Psychology, 10*, 173–220. https://doi.org/10.1016/S0065-2601(08)60357-3.
Rubin, D. B. (2018). *Multiple imputation*. Chapman and Hall/CRC.
Sabnis, S., Castillo, J. M., & Wolgemuth, J. R. (2020). RTI, equity, and the return to the status quo: Implications for consultants. *Journal of Educational and Psychological Consultation, 30*(3), 285–313. https://doi.org/10.1080/10474412.2019.1674152.
Silberglitt, B., & Hintze, J. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment, 23*(4), 304–325. https://doi.org/10.1080/02796015.2005.12086292.
Silva, M. R., Collier-Meek, M. A., Codding, R. S., Kleinert, W. L., & Feinberg, A. (2021). Data collection and analysis in response-to-intervention: A survey of school psychologists. *Contemporary School Psychology, 25*, 554–571.
Sims, W. A., King, K. R., Preast, J. L., Burns, M. K., & Panameño, S. (2024). Are school-based problem-solving teams effective? A meta-analysis of student-and systems-level effects. *Journal of Educational and Psychological Consultation, 34*(2), 115–139. https://doi.org/10.1080/10474412.2023.2232785.
Szadokierski, I., Burns, M. K., & McComas, J. J. (2017). Predicting intervention effectiveness from reading accuracy and rate measures through the instructional hierarchy: Evidence for a skill-by-treatment interaction. *School Psychology Review, 46*(2), 190–200.
Tarr, J. E., Reys, R. E., Reys, B. J., Chávez, Ó., Shih, J., & Osterlind, S. J. (2008). The impact of middle-grades mathematics curricula and the classroom learning environment on student achievement. *Journal for Research in Mathematics Education, 39*(3), 247–280.
Thiele, C., & Hirschfeld, G. (2020). *Cutpointr: Improved estimation and validation of optimal cutpoints in R*. arXiv preprint arXiv:2002.09209.
Thomas, A. S., & January, S. A. A. (2021). Evaluating the criterion validity and classification accuracy of universal screening measures in reading. *Assessment for Effective Intervention, 46*(2), 110–120. https://doi.org/10.1177/1534508419857231.
Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*, 1–67.
Van Norman, E. R., Klingbeil, D. A., & Nelson, P. M. (2017). Posttest probabilities: An empirical demonstration of their use in evaluating the performance of universal screening measures across settings. *School Psychology Review, 46*(4), 349–362. https://doi.org/10.17105/SPR-2017-0046.V46-4.
Van Norman, E. R., Nelson, P. M., & Klingbeil, D. A. (2017). Single measure and gated screening approaches for identifying students at-risk for academic problems: Implications for sensitivity and specificity. *School Psychology Quarterly, 32*(3), 405–413. https://doi.org/10.1037/spq0000177.
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.

VanMeveren, K., Hulac, D., & Wollersheim-Shervey, S. (2020). Universal screening methods and models: Diagnostic accuracy of reading assessments. *Assessment for Effective Intervention, 45*(4), 255–265.
Wang, M. C., Haertel, G. D., & Walberg, H. J. (1993). Toward a knowledge base for school learning. *Review of Educational Research, 63*(3), 249–294.
Ysseldyke, J. E., Chaparro, E. A., & VanDerHeyden, A. M. (2023). *Assessment in special and inclusive education*. Pro-Ed.