SMU | Annette Caldwell Simmons School of Education & Human Development

Research in Mathematics Education

Imagination Station (Istation): Istation's Indicators of Progress (ISIP) Math Validity Studies – Overview of Results

Technical Report 16-06

Imagination Station (Istation):
Istation's Indicators of Progress (ISIP) Math Validity Studies – Overview of Results

Pooja Shivraj • Lindsey Perry • Mo Zhao • Jason Bell • Leanne Ketterlin-Geller

Southern Methodist University

Fall 2016

Published by
Southern Methodist University
Department of Education Policy & Leadership
Simmons School of Education & Human Development
PO Box 750114
Dallas, TX 75275-0114
Contact information: rme@smu.edu

This research was supported by Imagination Station Inc. Opinions expressed herein do not necessarily reflect those of Imagination Station Inc. or individuals within.

Abstract

This report describes the evidence gathered to evaluate the appropriateness of Istation's Indicators of Progress (ISIP) Math for making screening decisions for students in Kindergarten through Grade 8. Evidence for the technical adequacy of ISIP Math for making screening decisions was collected to help Istation provide educators reasonable confidence in the inferences they make when using the ISIP Math data. Evidence gathered includes (a) generalizability of the sample, (b) classification accuracy of the performance level, (c) reliability of the scaled scores, (d) evidence for validity, and (e) evidence for reliability and validity disaggregated by relevant subgroup. Data for this study was obtained from three school districts in Texas during the 2015-2016 school year. Participants included eight schools and 108 teachers. A total of 2,038 students received parental consent and assented to participate in the study.

Overall, the evidence gathered suggests that the generalizability and reliability of ISIP Math within this study is moderate to strong across all grade levels. More evidence needs to be gathered for the technical adequacy of the Kindergarten ISIP Math using another criterion assessment with larger sample sizes. There is conflicting evidence presented for Grades 1 and 2, and more evidence needs to be gathered for these grades to determine the technical adequacy of Grades 1 and 2 ISIP Math. There is evidence for convincing classification within levels of "at-risk" and "not-at-risk" for Grades 3 through 6. There is also sufficient evidence for validity at these grade levels. Across all administrations, Grades 7 and 8 do not provide sufficient evidence of classification accuracy or validity. Coefficients disaggregated by relevant subgroup are also unstable in many cases. Additional research is needed to substantiate these results.

Table of Contents

Imagination Station (Istation): Istation's Indicators of Progress (ISIP) Math Validity Studies – Overview of Results

Introduction

Results from universal screening tools help teachers identify students who are on track and not on track for reaching their learning goals; screening tools can also be used to determine the intensity of instructional support that students may need to reach their curricular expectations by the end of the school year (Glover & Albers, 2007). In order to support educators in their instructional decision-making process, providing them with appropriate student assessment data that is substantiated by multiple sources of relevant evidence is important (AERA, APA, & NCME, 2014).

Reliability and validity are two sources of evidence commonly used to evaluate tests (AERA, APA, & NCME, 2014). Reliability typically refers to the consistency of measurement, while validity refers to the degree that interpretations made using test scores are appropriate, meaningful, and useful (Downing & Haladyna, 1997). More specific criteria have been put forth by the Technical Review Committee (TRC) from the National Center on Response to Intervention (NCRTI, 2010) for evaluating the technical adequacy of universal screening assessments. These include: (a) generalizability of the sample, (b) classification accuracy of the performance level, (c) reliability (of either the data or administrations of the assessment over time), (d) evidence for validity, and (e) evidence for reliability and validity disaggregated by relevant subgroup. The purpose of this study was to determine the appropriateness or technical adequacy of ISIP Math for making screening decisions for students in Kindergarten through Grade 8 based on these specified criteria. This report describes the technical adequacy data collected to document the utility of Istation's Indicators of Progress (ISIP) Math in making screening decisions for students in Kindergarten through Grade 8, and provides the overview of the results from this study. For a detailed description of the all components of this study, please refer to Shivraj et al. (2016).

Methods

Instruments

Istation's Indicators of Progress (ISIP) Math

ISIP Math is a web based computer adaptive formative assessment intended for students in Kindergarten through Grade 8. The assessment is intended to provide teachers and administrators with student test results to answer two questions: (a) whether students are at risk of failure, and (b) the degree of intensity of instructional support students need to be successful.

ISIP Math utilizes two unique interfaces for students depending on their grade level. Kindergarten and Grade 1 students are presented with a more interactive interface that incorporates drag and drop, tapping, and keypad entry while being supported by audio instructions. Grades 2-8 are presented with a single interface. Items are presented as multiple-choice and students must select one answer from four response options.

Kindergarten and Grade 1 content includes number sense, operations, geometry, algebra and algebraic thinking, measurement, data analysis, probability and statistics, personal financial literacy, and mathematical reasoning. Grades 2-8 content incorporates number and operations, geometry, algebra, measurement, ratios and proportional relationships, probability and statistics, personal financial literacy, and mathematical reasoning.

Four levels of cognitive engagement were also embedded into the assessment. The cognitive engagement refers to the level of cognitive processing with which students are expected to engage with the mathematical content. These cognitive engagement or demand levels were identified by the NRC (2001) and include procedural fluency, conceptual understanding, strategic competence, and adaptive reasoning. For details on the development and description of the content, please refer to Hatfield et al. (2015a, 2015b).

ISIP Math can be administered individually or in group settings. The assessment is administered in English. The assessment is untimed; however, most students complete the assessment within 30 minutes. There is no additional scoring time required for the assessment.

Teachers can be trained on ISIP Math through either a webinar or an in-person training session. Training takes between 1 and 4 hours. All training materials are online and are created by Istation.

There are three alternate forms of ISIP Math available per grade so that it can be administered up to three times during the school year. Reports are available for both individual and groups of students indicating single administration results and comparisons of results over time. All reports include student scaled scores and tier levels based on student percentiles.

Assessments Used to Obtain Criterion-Related Evidence of Validity for the ISIP Math

STAR Math

Renaissance Learning's STAR Math is also a computer adaptive test intended for students in Grades 1 through 8 (Renaissance Learning, 2015). STAR Math is designed to provide teachers and administrators with data that can be used for multiple purposes such as screening, placement, planning instruction, benchmarking, and outcomes measurement. It also provides educators with estimates of students' instructional math levels relative to national norms.

The assessment can be administered in groups or individually (Renaissance Learning, 2015). Students are given a three-minute time limit to complete individual items. Renaissance Learning indicates it takes an average time of 20 minutes for students to complete 34 items. Content includes numbers and operations, algebra, geometry and measurement, and data analysis, statistics, and probability. Reports are available for both individual and groups of students

indicating single administration results and over-time comparisons. Reports include students' scale scores, grade equivalent, percentile rank, normal equivalent score, and students' growth percentages. Internal consistency reliabilities ranged from .90 - .95 across grades, with the test-retest coefficient ranging from .76 - .84. Predictive and concurrent correlations ranged from moderate to strong, with predictive correlations ranging from r = .63 -. 80, and concurrent correlations ranging from r = .57 - .68.

Because STAR Math assesses a similar construct and has a similar use, STAR Math was used to provide criterion-related evidence for ISIP Math. However, it was not used as a criterion assessment or benchmark. Rather, the same set of analyses was conducted with both ISIP Math and STAR Math to examine similarities and differences in validity evidence. It was hypothesized that validity evidence would be comparable for these assessments and that this would provide strong criterion-related evidence for validity.

Test of Early Mathematics Ability –Third Edition (TEMA-3)

The Test of Early Mathematics Ability –Third Edition (TEMA-3) was selected as a criterion assessment for students in Kindergarten through Grade 2 for this study. TEMA-3 is intended to identify children who are significantly behind or ahead of their peers in the development of mathematical thinking (Ginsburg & Baroody, 2003). It is given to children age 4 to age 8 years 11 months. These students are typically associated with Kindergarten through Grade 2. The TEMA-3 is intended to be administered at the beginning of the school year as an early indicator of students' abilities, but can also be administered later in the school year to assess student progress.

The assessment is individually administered to the student and includes the use of manipulatives, a picture book, and student worksheet. A trained test administrator or data collector is required to manually enters student responses. The assessment is untimed and students may complete it in multiple sessions.

Mathematical concepts and skills assessed in the TEMA-3 include relative magnitude, counting, calculation, convention, number facts, base 10 concepts, non-verbal production, non-verbal addition and subtraction, part whole concepts, equal partitioning, symbolic additive commutativity, number comparisons, and mental addition and subtraction (Ginsburg & Baroody, 2003).

Reports include student raw score, percentile ranges, age equivalence, grade equivalence, and standard score. The TEMA-3 is available in two parallel forms, Form A and Form B. Research indicates that internal consistency reliabilities for both forms are above .92. Test-retest estimates are .82 for Form A and .93 for Form B. Ginsburg and Baroody (2003) also found that items in Form A contained bias. Given these findings, Form B was selected for this study. Criterion validity coefficients ranged from r = .36 - .71, with the majority of coefficients in the r = .50 - .60 range.

TEMA-3 was used as a criterion assessment, or as a benchmark, to support the inferences made from ISIP Math for Kindergarten through Grade 2.

Pearson's Stanford 10 (SAT10)

SAT10 was selected as a criterion assessment for student in grades 3-8 for this study. SAT10 online math assessment is a web-based multiple-choice assessment. The assessment is composed of two subtests, Mathematics Problem Solving (PS) and Mathematics Procedures (P), and contain 30 items and 20 items, respectively (Pearson, 2003). The assessment is proctored by a trained test administrator or data collector in a group setting. The assessment is untimed, but students are required to complete the assessment within 24 hours of starting. SAT10 is designed to assess mathematical content recommended by the National Council of Teachers of Mathematics Principles and Standards (2000). Student reports for both subtests include scaled score, percentile rank, normal curve equivalent, number correct, and stanine. Internal consistencies range from .80 - .87. Convergent validity coefficients range from r = .70 - .80 across grade levels.

SAT10 was used as a criterion assessment, or as a benchmark, to support the inferences made from ISIP Math for Grades 3 through 8.

State of Texas Assessments for Academic Readiness (STAAR)

State of Texas Assessments for Academic Readiness (STAAR) is the current state sponsored testing program in Texas (Texas Education Agency [TEA], 2013). The mathematics STAAR is a mandatory state assessment administered at the end of each school year between the months of March and May to students in Grades 3 through 8. It was first implemented in the 2011–2012 school year and replaced the previous mathematics state assessment, the Texas Assessment of Knowledge and Skills (TAKS).

Items on the STAAR are presented in a multiple-choice format with four available answers (TEA, 2013). Students are given 4 hours to complete each STAAR test. The assessment can be administered online or on paper. Students who take the mathematics STAAR receive a score for each of the four subdomains, a raw score, a scaled score, and one of three categories of proficiency (i.e., Advanced Academic Performance, Satisfactory Academic Performance, Unsatisfactory). STAAR is designed to measure the extent to which students are able to apply the knowledge and skills defined in the state curriculum standards. Internal consistency reliabilities for STAAR range from .81 - .93 across grade levels.

STAAR was also used as a criterion assessment, or as a benchmark, to support the inferences made from ISIP Math for Grades 3 through 8.

Instrument Administration Timeline

ISIP Math and STAR Math were delivered at the beginning of the year (BOY), the middle of the year (MOY), and the end of the year (EOY). TEMA-3, SAT10 (and its two subtests), and STAAR were only delivered at the EOY.

Sample

Data for this study was obtained from three school districts in Texas during the 2015-2016 school year. Participants included eight schools and 108 teachers. A total of 2,038 students received parental consent and assented to participate in the study. There were 178 kindergartners, 239 first graders, 218 second graders, 210 third graders, 137 fourth grader, 254 fifth graders, 279 sixth graders, 292 seventh graders, and 231 eighth graders. Table 1 shows the demographics of the recruited sample. It is important to note that not all students took all assessments at all testing administration windows.

Table 1
Demographics of the recruited sample

Demographics	Sample Distribution
Race	
Black or African American	13.76%
Hispanic	36.05%
White	42.88%
American Indian or Alaskan Native	0.42%
Asian	4.83%
Native Hawaiian/Other or Pacific Islander	0.42%
Two or More Races	2.23%
Gender	
Male	51.29%
Female	48.71%
Free/Reduced Lunch	
Yes	47.86%
No	52.14%

Table 2 shows the number of students participating in each of these assessments by grade. This only includes students who completed each of these assessments and who were included in the analysis.

Table 2
Number of students assessed by grade and assessment

Grade	ISIP Math			STAR Math			TEMA-3	SAT10	STAAR
	BOY	MOY	EOY	BOY	MOY	EOY	(EOY)	(EOY)	(EOY)
K	171	170	163				156		
1	217	230	221	222	214	216	215		
2	217	211	207	186	186	182	195		
3	208	204	198	171	171	171		201	196
4	133	134	133	83	71	81		131	129
5	248	251	252	228	198	235		250	241
6	247	260	250	178	173	167		202	254

7	268	250	237	244	215	239		162	229
8	189	196	201	195	192	157		163	147

Analyses

The analyses conducted align with the criteria identified by the NCRTI (2010): (a) generalizability of the sample, (b) classification accuracy of the performance level, (c) reliability (of either the data or administrations of the assessment over time), (d) evidence for validity, and (e) evidence for reliability and validity disaggregated by relevant subgroup. Details on the analyses performed for each of the criteria are detailed below. All analyses were performed in R (R Core Team, 2016).

Generalizability of the sample: To determine generalizability, the sample demographic characteristics were compared to the overall state demographic characteristics.

Classification accuracy of the performance level: Classification accuracy (also known as conditional probability analyses) refers to the extent to which an assessment is able to accurately differentiate between categories of students (Glover & Albers, 2007). Classification accuracy of the end-of-year ISIP Math (K-8) was calculated using the three criterion assessments: TEMA (K-2), SAT10 (Grades 3-8), and STAAR (Grades 3-8). For these analyses, the following criteria was used to classify students' scores within each assessment:

- ISIP Math: At-risk/Not-at-risk; categories based on percentile ranks, as determined by Istation. Tiers 2 and 3 were considered "at-risk." Students identified as needing Tier 1 instructional support were considered "not-at-risk."

- TEMA-3: At-risk/Not-at-risk; categories based on percentile ranks, calculated from ability score (Ginsburg & Baroody, 2003). Students who perform at or below the 20th percentile were classified as "at-risk." Students scoring at or above the 21st percentile are considered "not-at-risk."

- SAT10: Below Average/At or Above Average; categories based on stanines as described in the Technical Manual (Pearson, 2003)

    - SAT10 had scaled scores for the whole test (Total), the Mathematics Procedures (P) subtest, and the Mathematics Problem Solving (PS) subtest. SAT10 and both its subtests were used for analyses. The categories used for analyses were consistent across these subtests.

- STAAR: Satisfactory/Not Satisfactory; categories based on percent scores determined by the state.

For each of these assessments, the following statistics were calculated (described in no order of precedence):

- False positive rate: The proportion of "not-at-risk" students incorrectly identified as "at-risk". This is also known as a Type I Error. For example, the proportion of students identified by the ISIP Math as "at-risk" who were identified as "not-at-risk" on the TEMA-3 (or any of the other criterion measures), hence an incorrect identification of "at-risk" with respect to the criterion measure.

- False negative rate: The proportion of "at-risk" students incorrectly identified as "not-at-risk". This is also known as a Type II Error. For example, the proportion of students identified by the ISIP Math as "not-at-risk" who were identified as "at-risk" on the TEMA-3 (or any of the other criterion measures), hence an incorrect identification of "not-at-risk" with respect to the criterion measure.

- Sensitivity: The proportion of "at-risk" students correctly identified as "at-risk". This is also the true positive rate. For example, the proportion of students identified by the ISIP Math as "at-risk" who were identified as "at-risk" on the TEMA-3 (or any of the other criterion measures), hence a correct identification of "at-risk" with respect to the criterion measure.

- Specificity: The proportion of "not-at-risk" students correctly identified as "not-at-risk". This is also the true negative rate. For example, the proportion of students identified by the ISIP Math as "not-at-risk" who were identified as "not-at-risk" on the TEMA-3 (or any of the other criterion measures), hence a correct identification of "not-at-risk" with respect to the criterion measure.

- Positive predictive value: This is known as the precision. It is the proportion of students that are truly "at-risk" of all those identified as "at-risk". For example, of all the students identified as "at-risk" on both the ISIP Math and the TEMA-3 (or any of the other criterion measures), the positive predictive value is the proportion of students that are identified as "at-risk" on the TEMA-3 (or the criterion measure).

- Negative predictive value: The proportion of students that are truly "not-at-risk" of all those identified as "not-at-risk". For example, of all the students identified as "not-at-risk" on both the ISIP Math and the TEMA-3 (or any of the other criterion measures), the negative predictive value is the proportion of students that are identified as "not-at-risk" on the TEMA-3 (or the criterion measure).

- Accuracy: The proportion of correctly identified "at-risk" and "not-at-risk" students. For example, the accuracy of the ISIP Math reflects the proportion of students who were correctly identified as both "at-risk" and "not-at-risk" with respect to the criterion measure.

- Area Under the Curve (AUC): The probability that an assessment will predict "at-risk" when the true classification is "not-at-risk." Kettler et al. (2014) notes that AUC indices equal to or exceeding .80 are considered high and indicative of strong universal screening systems. Indices greater than .60 but less than .80 are considered moderate. However, the guidelines for the NCRTI (2010) state that AUC indices of greater than .85 are considered to have convincing evidence. AUC indices of .75-.85 are considered to have

partially convincing evidence, while AUC indices of less than .75 are considered to have unconvincing evidence.

Reliability: Since the ISIP Math is a computerized adaptive test and students took different sets of items, a sparse data matrix was obtained. Because of this, traditional reliability analysis (i.e., Cronbach's alpha) could not be conducted. Instead, reliability was calculated using the standard error of measurement. The standard errors, which are conditional on the scaled scores, were pre-calculated as the standard deviation of the sampling distribution of the estimated ability at students' true ability. The following steps were then taken to calculate reliability (Bechger, Maris, Verstralen, & Beguin, 2003):

- The conditional error variance was calculated as the square of the obtained standard error of measurement for the scaled scores for each student.
- The error score variance was calculated to be the mean of the conditional error variance.
- The observed (or estimated) score variance was calculated to be the variance of the estimated abilities.
- The true score variance was calculated to be the difference between the observed score variance and the error score variance.
- The reliability was obtained by taking the ratio of the true score variance to the observed score variance. This reliability estimate indicates how consistent the scale scores obtained are, or given resampling, whether a similar estimated ability for a student's true ability would be obtained.

The guidelines for the NCRTI (2010) state that two types of reliability indices (split half, alpha, test-retest, or inter-rater) with a coefficient of greater than .80 have to be presented to provide *Convincing Evidence*. If only one is presented that is greater than .80, there exists *Partially Convincing Evidence*.

Evidence for validity: Criterion-related evidence for validity should be considered when evaluating the technical adequacy of screeners. This type of evidence serves as an indicator of the extent to which a test taker's performance on the universal screener is associated with that same person's performance on a criterion measure, such as a norm-referenced test or a state accountability test (Johnson, Jenkins, Petscher, & Catts, 2009). Predictive-related evidence for validity examines the relation between performance on the screener and a criterion of similar content that is administered at some time in the future; in contrast, concurrent-related evidence for validity examines the relation between performance on the screener and a criterion of similar content that is administered at the same point in time. Kline (2000) suggests that coefficients of about .75 serve as strong indicators of evidence for concurrent validity, correlations of .4 to .5 serve as moderate indicators of evidence for concurrent validity, and correlations of .3 to .4 serve as moderate indicators of evidence for predictive validity.

Both concurrent-related and predictive-related evidence for validity were collected. Concurrent-related evidence for validity was collected relative to STAR Math for all three administrations of

ISIP Math – BOY ISIP Math to BOY STAR Math, MOY ISIP Math to MOY STAR Math, and EOY ISIP Math to EOY STAR Math. Concurrent-related evidence for validity was collected relative to the three criterion assessments that were delivered at the end of the year – TEMA-3, SAT10 (along with both its subtests - Mathematics Problem Solving subtest [PS] and Mathematics Procedure subtest [P]), and STAAR – for the EOY ISIP Math. Predictive-related evidence for validity for the BOY and MOY administrations of ISIP Math were also collected relative to the three criterion assessments. The guidelines for the NCRTI (2010) state that correlation indices for validity of greater than .70 are considered to have *Convincing Evidence*. The NCRTI guidelines also suggest collective evidence for coefficients by relevant subgroup. Each of the concurrent and validity analyses were disaggregated by gender, race, and economic disadvantage.

Evidence for reliability and validity disaggregated by relevant subgroup: Each of the reliability and validity coefficients were disaggregated by gender (male/female), economically disadvantaged (yes/no), and race (White/African American/Hispanic). Other races were not included due to limited sample sizes. Eligibility for free and reduced priced meals in the National School Lunch Program was used as a proxy for the economically disadvantaged variable.

Results and Discussion

Technical adequacy data were collected to document the utility of ISIP Math in making screening decisions for students in Kindergarten through Grade 8. The criteria used within this study were identified by the NCRTI (2010); these include: (a) generalizability of the sample, (b) classification accuracy of the performance level, (c) reliability (of either the data or administrations of the assessment over time), (d) evidence for validity, and (e) evidence for reliability and validity disaggregated by relevant subgroup. Results from this study are presented next.

Generalizability

Generalizability was analyzed as a way to illustrate the extent to which the analytic sample for the study was comparable to the state and national population. The data in Table 3 shows a comparison of the demographics for the state (2015-16), national (2012-14), and the recruited sample.

Table 3
Comparison of demographics for the state (2015-16) and recruited sample

Demographics	Statewide Distribution[a]	National Distribution[bcd]	Sample Distribution
Race/Ethnicity			
Black or African American	12.61%	15.60%	13.76%
Hispanic	52.22%	24.88%	36.05%
White	28.55%	50.28%	42.88%
American Indian or Alaskan Native	0.39%	1.05%	0.42%
Asian	4.03%	5.18%	4.83%
Native Hawaiian/Other or Pacific Islander	0.14%		0.42%

Two or More Races	2.05%	3.02%	2.23%
Gender			
Male	51.30%	51.4%	51.29%
Female	48.70%	48.6%	48.71%
Free/Reduced Lunch			
Yes	50.10%	48.1%	47.86%
No	49.90%	51.9%	52.14%

[a]Texas Education Agency (2015). [b]U.S. Department of Education, National Center for Education Statistics, and Common Core of Data (2012). [c]U.S. Department of Education, National Center for Education Statistics, and Common Core of Data (2016). [d]U.S. Census Bureau (2014).

The data in Table 3 indicate that while the sample is comparable to the national and state population with respect to economically disadvantaged status and state population of gender, the percent of students in different racial groups vary. The sample is comparable to the state population with respect to Black/African American, Two or More Races, and other ethnicity/races (i.e., American Indian or Alaskan Native, Asian, Native Hawaiian/Other or Pacific Islander); however, the sample has a lower Hispanic and higher White composition than the state population (36.05% compared to 52.22% statewide, and 42.88% compared to 28.55% statewide, respectively). Conversely, it has a higher Hispanic and lower White composition than the national population (36.05% compared to 24.88% nationally, and 42.88% compared to 50.28% nationally, respectively). Thus, the sample is close to mirroring both state and national representation, apart from the White and Hispanic composition of the sample.

Classification Accuracy

Classification accuracy analyses were performed to determine if ISIP Math was able to accurately differentiate between categories of students ("at-risk" vs. "not-at-risk") using TEMA-3, SAT10, and STAAR as the criterion assessments.

Classification Accuracy for Grades K-2

Tables 4-9 provide the classification accuracy values for the EOY ISIP Math with respect to the TEMA-3, SAT10 (and its two subtests), and the STAAR. In the following section, sensitivity and specificity are discussed, while the False Positive Rate (FPR) and the False Negative Rate (FNR) are found in the tables but not specifically discussed. This is because the FPR is the additive inverse of the specificity and FNR is the additive inverse of the sensitivity. The PPV and NPV are discussed along with the accuracy and AUC.

First, Table 4 provides the classification accuracy values for the EOY ISIP Math with respect to TEMA-3. Classification accuracy values for Kindergarten through Grade 2 are provided since TEMA-3 is administered to students at those grade levels only.

Table 4
Classification accuracy of EOY ISIP Math on TEMA-3

Grade	n	FPR	FNR	Sens.	Spec.	PPV	NPV	Acc.	AUC
K	152	.21	.20	.80	.79	.97	.29	.80	.80
1	210	.39	.13	.87	.61	.90	.53	.82	.74
2	195	.23	.08	.92	.77	.94	.70	.89	.84

Note. FPR: False Positive Rate; FNR: False Negative Rate; Sens: Sensitivity; Spec: Specificity; PPV: Positive Predictive Value; NPV: Negative Predictive Value; Acc: Accuracy; AUC: Area Under the Curve

From Table 4, the sensitivity of ISIP Math across Kindergarten through Grade 2 using TEMA-3 as the criterion assessment was between .80 and .92. In other words, between 80% and 92% of the students who were classified as "at-risk" on the TEMA-3 were also classified as "at-risk" on the EOY ISIP Math.

The specificity of ISIP Math across Kindergarten through Grade 2 using TEMA-3 as the criterion assessment was lower, between .61 and .79 across grades. In other words, between 61% and 79% of the students who were classified as "not-at-risk" on the TEMA-3 were also classified as "not-at-risk" on the EOY ISIP Math. This also indicates that 21-39% of students were classified as "at-risk" on the ISIP Math were classified as "not-at-risk" on the TEMA-3.

The PPV, or precision of classification, range from .90-.97 across grades. This indicates that 90-97% of the students who were truly "at-risk" were classified as "at-risk" on both the ISIP Math and the TEMA-3. The NPV ranges from .29-.70 across grades, indicating that 29-70% of students who were truly "not-at-risk" were classified as "not-at-risk" on both the ISIP Math and the TEMA-3. The NPV value coincides with the specificity in that a large proportion of students who were classified as "at-risk" on the EOY ISIP Math were classified as "not-at-risk" on the TEMA-3.

The accuracy of identification ranges from .80 to .89, indicating that the percent of students correctly classified on the EOY ISIP Math with respect to the TEMA-3 was between 80% and 89% across all grades. The AUC indices range from .74 to .84 across grades. Using the guidelines suggested by Kettler et al. (2014), the AUC indices are moderate to high. Using the guidelines set by the NCRTI (2010), Kindergarten and Grade 2 ISIP Math provide *Partially Convincing Evidence* for classification accuracy based on TEMA-3, while Grade 1 ISIP Math provides *Unconvincing Evidence* for classification accuracy based on TEMA-3.

One possible explanation for over-classification of "at-risk" students is that the cut score used for classification of "at-risk" and "not-at-risk" on the TEMA-3 is the 20th percentile, while the cut score used for ISIP Math is the 25th percentile.

Classification Accuracy for Grades 3-8

Compared to SAT10

Tables 5-7 provide the classification accuracy values for the EOY ISIP Math with respect to SAT10 and its two subtests (Mathematics Procedures and Mathematics Problem Solving).

Table 5
Classification accuracy of EOY ISIP Math on SAT10

Grade	n	FPR	FNR	Sens.	Spec.	PPV	NPV	Acc.	AUC
3	196	.32	.05	.95	.68	.95	.70	.91	.81
4	131	.18	.12	.88	.82	.97	.50	.87	.85
5	250	.32	.02	.98	.68	.97	.75	.95	.83
6	197	.48	.03	.97	.52	.89	.81	.88	.75
7	146	.58	.04	.96	.42	.85	.74	.84	.69
8	152	.75	.05	.95	.25	.98	.11	.93	.60

Note. FPR: False Positive Rate; FNR: False Negative Rate; Sens: Sensitivity; Spec: Specificity; PPV: Positive Predictive Value; NPV: Negative Predictive Value; Acc: Accuracy; AUC: Area Under the Curve

From Table 5, the sensitivity of ISIP Math across Grades 3 through 8 using SAT10 as the criterion assessment was between .88 and .98. In other words, between 88% and 98% of the students who were classified as "at-risk" on SAT10 were also classified as "at-risk" on the EOY ISIP Math.

The specificity of ISIP Math across Grades 3 through 8 using SAT10 as the criterion assessment was lower, between .25 and .82. In other words, between 25% and 82% of the students who were classified as "not-at-risk" on SAT10 were also classified as "not-at-risk" on the EOY ISIP Math. This also suggests that 18-75% of students who were classified as "at-risk" on ISIP Math were classified as "not-at-risk" on SAT10.

The PPV, or precision of classification, range from .85-.98 across grades. This indicates that 85-98% of the students who were truly "at-risk" were classified as "at-risk" on both the ISIP Math and SAT10. The NPV ranges from .11-.81 across grades, indicating that 11-81% of students who were truly "not-at-risk" were classified as "not-at-risk" on both the ISIP Math and SAT10. The NPV value coincides with the specificity in that a large proportion of students classified as "at-risk" on the EOY ISIP Math were classified as "not-at-risk" on SAT10.

The accuracy of identification ranges from .84 to .95, indicating that the percent of students correctly classified on the EOY ISIP Math with respect to SAT10 was between 84% and 95% across all grades. The AUC indices range from .60 to .85 across grades. Using the guidelines suggested by Kettler et al. (2014), the AUC indices are moderate to high. Using the guidelines set by the NCRTI (2010), Grades 3 through 6 on the ISIP Math provide *Partially Convincing Evidence* for classification accuracy based on SAT10, while Grades 7 and 8 ISIP Math provides *Unconvincing Evidence* for classification accuracy based on SAT10.

The differences in classification on the ISIP Math and SAT10 may be a function of differences in the cut scores used by each test. Another plausible explanation is that SAT10 comprises two main cognitive engagement levels – procedures and problem solving – while ISIP Math comprises four main cognitive engagement levels – procedural fluency, conceptual understanding, strategic competence, and adaptive reasoning.

Table 6 provides the classification accuracy values for the EOY ISIP Math with respect to the SAT10 Mathematics Procedures subtest.

Table 6

Classification accuracy of EOY ISIP Math on the SAT10 Procedures subtest

Grade	n	FPR	FNR	Sens.	Spec.	PPV	NPV	Acc.	AUC
3	196	.50	.04	.96	.50	.88	.78	.86	.73
4	131	.41	.11	.88	.59	.89	.57	.82	.74
5	250	.64	.01	.99	.36	.86	.90	.86	.67
6	197	.57	.03	.97	.43	.83	.85	.83	.70
7	146	.69	.05	.95	.31	.76	.74	.75	.63
8	152	.67	.05	.95	.33	.99	.11	.93	.64

Note. FPR: False Positive Rate; FNR: False Negative Rate; Sens: Sensitivity; Spec: Specificity; PPV: Positive Predictive Value; NPV: Negative Predictive Value; Acc: Accuracy; AUC: Area Under the Curve

From Table 6, the sensitivity of ISIP Math across Grades 3 through 8 using SAT10 Mathematics Procedures subtest as the criterion assessment was between .88 and .99. In other words, between 88% and 99% of the students who were classified as "at-risk" on the SAT10 Mathematics Procedures subtest were also classified as "at-risk" on the EOY ISIP Math.

The specificity of ISIP Math across Grades 3 through 8 using the SAT10 Mathematics Procedures subtest as the criterion assessment was lower, between .31 and .59. In other words, between 31% and 59% of the students who were classified as "not-at-risk" on the SAT10 Mathematics Procedures subtest were also classified as "not-at-risk" on the EOY ISIP Math. This also suggests that 41-69% of students who were classified as "at-risk" were classified as "not-at-risk" on the SAT10 Mathematics Procedures subtest.

The PPV, or precision of classification, ranged from .76-.99 across grades. This indicates that 76-99% of the students who were truly "at-risk" were classified as "at-risk" on both the ISIP Math and the SAT10 Mathematics Procedures subtest. The NPV ranged from .11-.90 across grades, indicating that 11-90% of students who were truly "not-at-risk" were classified as "not-at-risk" on both the ISIP Math and the SAT10 Mathematics Procedures subtest. The NPV value coincides with the specificity in that a large proportion of students classified as "at-risk" by the EOY ISIP Math were classified as "not-at-risk" on the SAT10 Mathematics Procedures subtest.

The accuracy of identification ranges from .75 to .93, indicating that the percent of students who were correctly classified on the EOY ISIP Math with respect to the SAT10 Mathematics Procedures subtest was between 75% and 93% across all grades. The AUC indices range from .63 to .74 across grades. Using the guidelines suggested by Kettler et al. (2014), the AUC indices are moderate. Using the guidelines set by the NCRTI (2010), all grades on the ISIP Math provide *Unconvincing Evidence* for classification accuracy based on the SAT10 Mathematics Procedures subtest.

As suggested earlier, a plausible explanation for the differences in classification could be that the SAT10 Mathematics Procedures subtest and ISIP Math assess different cognitive engagement levels. The SAT10 Mathematics Procedures subtest comprises only procedural items, while ISIP Math comprises four main cognitive engagement levels – procedural fluency, conceptual understanding, strategic competence, and adaptive reasoning. Therefore, classification accuracy on the SAT10 Mathematics Procedures subtest would not reflect all cognitive competencies of the assessment.

Table 7 provides the classification accuracy values for the EOY ISIP Math with respect to the SAT10 Mathematics Problem Solving subtest.

Table 7
Classification accuracy of EOY ISIP Math on the SAT10 Problem Solving subtest

Grade	n	FPR	FNR	Sens.	Spec.	PPV	NPV	Acc.	AUC
3	196	.28	.05	.95	.72	.96	.67	.92	.83
4	131	.15	.14	.86	.85	.98	.39	.86	.85
5	250	.14	.06	.94	.86	1.00	.30	.94	.90
6	197	.41	.05	.95	.59	.93	.65	.89	.77
7	146	.63	.07	.93	.37	.85	.58	.81	.65
8	152	.79	.04	.96	.21	.92	.33	.89	.58

Note. FPR: False Positive Rate; FNR: False Negative Rate; Sens: Sensitivity; Spec: Specificity; PPV: Positive Predictive Value; NPV: Negative Predictive Value; Acc: Accuracy; AUC: Area Under the Curve

From Table 7, the sensitivity of ISIP Math across Grades 3 through 8 using the SAT10 Mathematics Problem Solving subtest as the criterion assessment was between .86 and .96 This indicates that between 86% and 96% of the students who were classified as "at-risk" on the SAT10 Mathematics Problem Solving subtest were also classified as "at-risk" on the EOY ISIP Math. The specificity of ISIP Math across Grades 3 through 8 using the SAT10 Mathematics Problem Solving subtest as the criterion assessment was between .21 and .86. In other words, between 21% and 86% of the students who were classified as "not-at-risk" on the SAT10 Mathematics Problem Solving subtest were also classified as "not-at-risk" on the EOY ISIP Math.

The PPV, or precision of classification, ranges from .85-1.00 across grades. This indicates that 85-100% of the students who were truly "at-risk" were classified as "at-risk" on both the ISIP Math and the SAT10 Mathematics Problem Solving subtest. The NPV ranged from .30-.67 across grades, indicating that 30-67% of students who were truly "not-at-risk" were identified as "not-at-risk" on both the ISIP Math and the SAT10 Mathematics Problem Solving subtest. The NPV value coincides with the specificity in that a large proportion of students classified as "at-risk" by the EOY ISIP Math were classified as "not-at-risk" on the SAT10 Mathematics Problem Solving subtest.

The accuracy of identification ranges from .81 to .94, indicating that the percent of students who were correctly classified on the EOY ISIP Math with respect to the SAT10 Mathematics Problem Solving subtest was between 81% and 94% across all grades. The AUC indices range from .58 to .90 across grades. Using the guidelines suggested by Kettler et al. (2014), the AUC indices are low to high. Using the guidelines set by the NCRTI (2010), Grades 3, 4, and 6 provide *Partially Convincing Evidence* based on the SAT10 Mathematics Problem Solving subtest; Grade 5 provides *Convincing Evidence*; and Grades 7 and 8 provide *Unconvincing Evidence* for classification accuracy based on the SAT10 Mathematics Problem Solving subtest.

Given that there is stronger evidence for classification accuracy based on the SAT10 Mathematics Problem Solving subtest for Grades 3-6, the content and cognitive engagement levels assessed on this subtest may be more closely aligned to the items on the ISIP Math, which may be further investigated.

Compared to STAAR

Table 8 provides the classification accuracy values for the EOY ISIP Math with respect to STAAR. Classification accuracy values are only provided for Grades 3 through 8 since the mathematics STAAR is a mandated state assessment for students at those grade levels.

Table 8
Classification accuracy of EOY ISIP Math on STAAR

Grade	n	FPR	FNR	Sens.	Spec.	PPV	NPV	Acc.	AUC
3	190	.21	.02	.98	.79	.97	.83	.95	.88
4	129	.31	.09	.91	.69	.92	.67	.87	.80
5	241	.12	.02	.98	.88	.99	.74	.97	.93
6	234	.53	.07	.93	.47	.92	.48	.87	.70
7	192	.50	.02	.98	.50	.98	.50	.97	.74
8	130	.80	.06	.94	.20	.97	.11	.91	.57

Note. FPR: False Positive Rate; FNR: False Negative Rate; Sens: Sensitivity; Spec: Specificity; PPV: Positive Predictive Value; NPV: Negative Predictive Value; Acc: Accuracy; AUC: Area Under the Curve

From Table 8, the sensitivity of ISIP Math across Grades 3 through 8 using STAAR as the criterion assessment was between .91 and .98. Therefore, between 91% and 98% of the students who were classified as "at-risk" on the STAAR were also classified as "at-risk" on the EOY ISIP Math. The specificity of ISIP Math across Grades 3 through 8 using STAAR as the criterion assessment was between .20 and .88. In other words, between 20% and 88% of the students who were classified as "not-at-risk" on the STAAR were also classified as "not-at-risk" on the EOY ISIP Math.

The PPV, or precision of classification, ranges from .92-.99 across grades. This indicates that 92-99% of the students who were truly "at-risk" were identified as "at-risk" on both the ISIP Math and the STAAR. The NPV ranges from .11-.83 across grades, indicating that 11-83% of students who were truly "not-at-risk" were classified as "not-at-risk" on both the ISIP Math and the STAAR. The NPV value coincides with the specificity in that a large proportion of students classified as "at-risk" by the EOY ISIP Math were classified as "not-at-risk" on the STAAR.

The accuracy of identification ranges from .87 to .97, indicating that the percent of students who were correctly identified on the EOY ISIP Math with respect to the STAAR was between 87% and 97% across all grades. The AUC indices range from .57 to .93 across grades. Using the guidelines suggested by Kettler et al. (2014), the AUC indices are low to high. Using the guidelines set by the NCRTI (2010), Grades 3 through 5 provide *Partially Convincing Evidence* (Grade 4) to *Convincing Evidence* (Grades 3 and 5) based on the STAAR; however, Grades 6 through 8 provide *Unconvincing Evidence* for classification accuracy based on the STAAR.

Plausible reasons for differences in classification between ISIP Math and STAAR include variability in the cut scores across assessments and a differential balance of items based on content and/or cognitive engagement levels. Another possible reason may be the change in standards in 2013-14, which may have impacted student scores in having them have less time in preparing for content that aligns with these standards. This needs to be further investigated. Additionally, student motivation may have impacted these scores since STAAR is a high-stakes

assessment and ISIP Math is a low-stakes assessment. Each of these possibilities may be further investigated.

Summary of Classification Accuracy for Grades K-8

Table 9 summarizes the evidence across all assessments and grade levels for the classification accuracy of the EOY ISIP Math based on the NCRTI (2010) criteria. The criteria NCRTI uses to determine level of evidence for classification accuracy is Area Under the Curve (AUC).

Table 9
Summary of evidence for classification accuracy using the NCRTI (2010) criteria for AUC

	TEMA-3	SAT10	SAT10 P	SAT10 PS	STAAR
K	Partially Convincing				
1	Unconvincing				
2	Partially Convincing				
3		Partially Convincing	Unconvincing	Partially Convincing	Convincing
4		Partially Convincing	Unconvincing	Partially Convincing	Partially Convincing
5		Partially Convincing	Unconvincing	Convincing	Convincing
6		Partially Convincing	Unconvincing	Partially Convincing	Unconvincing
7		Unconvincing	Unconvincing	Unconvincing	Unconvincing
8		Unconvincing	Unconvincing	Unconvincing	Unconvincing

There are three main conclusions and implications from the evidence provided in this section. First, ISIP Math generally classifies more students as "at-risk" than are actually "at-risk", as classified by the criterion assessment. This is known as a Type I error. For screeners, it is ideal to make a Type I error (classifies students who are "not-at-risk" as "at-risk" and provide them support) than make a Type II error (classify students who are "at-risk" as "not-at-risk" and miss an opportunity to provide support). As shown in Tables 4-8, the rate of a Type II error (FNR) for ISIP Math across all assessments and grade levels is low.

Second, ISIP Math uses a standard criterion for differentiating scores across tiers, which may differ from summative assessments that have specific cut scores, criteria, and performance levels; therefore, interpretations of classification accuracy results should be made with caution.

Third, it should be noted that since ISIP Math is given in the early spring as a formative assessment, some interventions may have been applied before the summative assessments were delivered, which could have impacted how students performed and how they were classified.

Reliability

Reliability was calculated using the standard error of measurement of the scaled score. The estimates obtained indicate the consistency of estimates of students' ability levels. Table 10 summarizes the reliability coefficients for ISIP Math for each grade level at each administration.

Table 10
Evidence for reliability of ISIP Math

Grade	BOY	MOY	EOY
K	.89	.92	.95

1	.92	.94	.94
2	.90	.92	.93
3	.90	.88	.94
4	.85	.88	.93
5	.92	.87	.92
6	.81	.85	.92
7	.85	.88	.93
8	.83	.88	.91

Across all grade levels and administrations, there is *Partially Convincing Evidence* for reliability of ISIP Math based on the NCRTI (2010) criteria. This is because only one reliability coefficient has been presented, and NCRTI (2010) requires presenting at least two types of reliability indices (split half, alpha, test-retest, or inter-rater); however, all calculated indices are greater than the cutoff of .80, indicating replicable scaled scores. Additional evidence for reliability is needed using other methods.

Criterion-Related Evidence for Validity

Concurrent-Related Evidence For Validity

Concurrent-related evidence for validity examines the relation between performance on the screener and a criterion assessment with similar content that is administered at the same point in time. Concurrent-related evidence for validity at each administration of the ISIP Math was calculated by determining the correlation between the scaled scores of ISIP Math for that administration and the scaled scores of the STAR Math for the same administration by grade level. Concurrent-related evidence for validity at the EOY administration of the ISIP Math was also calculated by determining the correlation between the scaled scores of the EOY ISIP Math and the scaled scores of the TEMA-3, SAT10 (and its two subtests), and the STAAR, individually, by grade level.

The correlation coefficients and sample sizes for concurrent-related evidence for ISIP Math are presented in Table 11.

Table 11
Concurrent-related evidence for validity

Assessment	Grade	*n*	Coefficient
STAR Math (BOY)	1	208	.66
	2	185	.76
	3	170	.71
	4	81	.64
	5	224	.55
	6	174	.74
	7	222	.61
	8	165	.61
STAR Math (MOY)	1	212	.77
	2	183	.81
	3	169	.75
	4	69	.67
	5	198	.71

	6	173	.77
	7	199	.60
	8	167	.59
STAR Math (EOY)	1	213	.72
	2	181	.75
	3	167	.74
	4	81	.78
	5	235	.76
	6	162	.80
	7	211	.76
	8	145	.61
TEMA-3	K	152	.49
	1	210	.66
	2	195	.69
SAT10	3	196	.82
	4	131	.82
	5	250	.82
	6	197	.83
	7	146	.57
	8	152	.67
SAT10 PS	3	196	.82
	4	131	.82
	5	250	.75
	6	197	.83
	7	146	.45
	8	152	.65
SAT10 P	3	196	.69
	4	131	.71
	5	250	.78
	6	197	.74
	7	146	.58
	8	152	.54
STAAR	3	190	.81
	4	129	.80
	5	241	.81
	6	234	.85
	7	192	.70
	8	130	.68

The NCRTI (2010) established a criterion of .70 for demonstrating acceptable concurrent associations with the criterion assessment. First, Table 12 summarizes the concurrent-related evidence for the ISIP Math (see Table 11) based on this criterion. Then, interpretations of the coefficients (see Table 11) and the evidence (see Table 12) are discussed. For the concurrent-related evidence obtained from STAR Math, if all three coefficients (BOY, MOY, and EOY) were above .70, then the evidence was deemed as *Convincing*. If two coefficients were above .70, then the evidence was considered *Partially Convincing*.

Table 12

Summary of concurrent-related evidence for validity based on the NCRTI (2010) criteria

	TEMA-3	SAT10	SAT10 P	SAT10 PS	STAAR	STAR Math
K	Unconvincing					
1	Unconvincing					Partially Convincing
2	Unconvincing					Convincing
3		Convincing	Unconvincing	Convincing	Convincing	Convincing
4		Convincing	Convincing	Convincing	Convincing	Unconvincing
5		Convincing	Convincing	Convincing	Convincing	Convincing
6		Convincing	Convincing	Convincing	Convincing	Convincing
7		Unconvincing	Unconvincing	Unconvincing	Unconvincing	Unconvincing
8		Unconvincing	Unconvincing	Unconvincing	Unconvincing	Unconvincing

Using the NCRTI (2010) criterion, TEMA-3 provides *Unconvincing Evidence* for making concurrent-related validity interpretations about ISIP Math at Kindergarten through Grade 2. This could be due to multiple reasons. TEMA-3 may not assess the same constructs that ISIP Math assesses at those grade levels. TEMA-3 also uses a different administration technique than ISIP Math (individually administered paper-based versus group administered computer-based), which could lead to rater bias on the TEMA-3 scores. However, this is unlikely due to the extensive training provided to raters.

Using the same criterion, SAT10 provides *Convincing Evidence* for making concurrent-related validity interpretations about ISIP Math across Grades 3 through 6 (except for Grade 3 on the SAT10 Mathematics Procedures subtest); however, it provides *Unconvincing Evidence* for making concurrent-related validity interpretations about ISIP Math in Grades 7 and 8. Similar to TEMA-3, likely reasons may include differences in tested content on the Grade 7 and 8 SAT10 and ISIP Math. This may lead to weak associations between the scores on these assessments. This may be investigated further, and content-related evidence for validity could be collected and examined.

STAAR provides *Convincing Evidence* for making concurrent-related validity interpretations about ISIP Math from Grades 3 through 6. *Unconvincing Evidence* is provided for Grades 7 and 8, which may be due to similar reasons stated above.

The evidence provided by STAR Math varied across grade levels. Grades 2, 3, 5, and 6 provided *Convincing Evidence* for making concurrent-related interpretations; Grades 7 and 8, similar to SAT10, provided *Unconvincing Evidence*. Grade 4 also provided *Unconvincing Evidence*. Finally, Grade 1 provided *Partially Convincing Evidence*, with its MOY and EOY coefficients being at or above .70, dissimilar to the results of TEMA-3.

Across assessments and grade levels, concurrent-related validity evidence for K-2 may need to be gathered with a different criterion assessment since the results from TEMA-3 and STAR Math are contradictory. The results from SAT10 and STAAR mostly align, wherein there is *Convincing Evidence* for Grades 3 through 6 but not necessarily for Grades 7 and 8. These grade levels may need a deeper analysis of the content, including an investigation of the blueprint and items, or may require an additional criterion assessment to confirm the results.

Predictive-Related Evidence For Validity

Predictive-related evidence for validity examines the relation between performance on the screener and a criterion assessment with similar content that is administered at some time in the future. Predictive-related evidence for validity for the BOY and MOY administrations of ISIP Math were collected relative to the EOY criterion assessments - TEMA-3, SAT10 (and both its subtests), and STAAR.

The correlation coefficient ranges and sample sizes for predictive-related evidence for BOY and MOY ISIP Math are presented in Table 13.

Table 13
Predictive-related evidence for validity

		BOY ISIP Math		MOY ISIP Math	
Assessment	Grade	*n*	Coefficient	*n*	Coefficient
TEMA-3	K	154	.51	153	.44
	1	199	.65	215	.74
	2	195	.60	193	.70
SAT10	3	200	.73	198	.71
	4	127	.71	129	.76
	5	244	.57	247	.75
	6	196	.72	197	.72
	7	146	.43	138	.51
	8	134	.62	139	.67
SAT10 Problem Solving	3	200	.75	198	.73
	4	127	.72	129	.75
	5	244	.58	247	.72
	6	196	.72	197	.73
	7	146	.31	138	.43
	8	134	.62	139	.64
SAT10 Procedures	3	200	.60	198	.58
	4	127	.58	129	.65
	5	244	.48	247	.67
	6	196	.63	197	.63
	7	146	.48	138	.48
	8	134	.50	139	.54
STAAR	3	195	.65	194	.68
	4	125	.70	127	.77
	5	235	.54	238	.74
	6	230	.72	239	.79
	7	212	.69	205	.66
	8	139	.57	140	.65

The NCRTI (2010) established a criterion of .70 for demonstrating acceptable predictive associations with the criterion assessment. First, Table 14 summarizes the evidence for predictive-related evidence for the ISIP Math based on this criterion. Then, interpretations of the coefficients (see Table 13) and the evidence (see Table 14) are discussed. For the predictive-

related evidence, if at least one coefficient (either the BOY or the MOY ISIP Math coefficient) was at or above .70, then the evidence was deemed as *Partially Convincing*. If both coefficients were at or above .70, then the evidence was considered *Convincing*.

Table 14
Summary of predictive-related evidence for validity based on the NCRTI (2010) criteria

	TEMA-3	SAT10	SAT10 P	SAT10 PS	STAAR
K	Unconvincing				
1	Partially Convincing				
2	Partially Convincing				
3		Convincing	Unconvincing	Convincing	Unconvincing
4		Convincing	Unconvincing	Convincing	Convincing
5		Partially Convincing	Unconvincing	Partially Convincing	Partially Convincing
6		Convincing	Unconvincing	Convincing	Convincing
7		Unconvincing	Unconvincing	Unconvincing	Unconvincing
8		Unconvincing	Unconvincing	Unconvincing	Unconvincing

Using the NCRTI (2010) criterion, TEMA-3 provides *Unconvincing Evidence* for making predictive-related validity interpretations about ISIP Math at the Kindergarten level but *Partially Convincing Evidence* at Grades 1 and 2 (with the MOY ISIP Math association with TEMA-3 being stronger). The reasons for lower associations may be similar or the same as the reasons for low concurrent-related correlations. The time of year when students learn the content may also play a role in how these correlations vary.

SAT10 and the SAT10 Mathematics Problem Solving subtest provide similar levels of evidence at all grade levels; it provides *Convincing Evidence* for making predictive-related validity interpretations about ISIP Math at Grades 3, 4, and 6, and *Partially Convincing Evidence* at Grade 5. However, both tests provide *Unconvincing Evidence* for making predictive-related validity interpretations about ISIP Math in Grades 7 and 8. The SAT10 Mathematics Procedures subtest also provides *Unconvincing Evidence* for predictive-related interpretations about ISIP Math at all grade levels, which may be due to an imbalance of cognitive levels on the BOY and MOY ISIP Math assessments or the criterion assessments. The results of the predictive-related evidence for Grades 7 and 8 align with the concurrent-related evidence.

Similar to both the validity evidence provided by SAT10 and the concurrent-related evidence for Grades 7 and 8 provided by STAAR, STAAR provides *Unconvincing Evidence* for making predictive-related validity interpretations about ISIP Math for Grades 7 and 8. However, *Partially Convincing* to *Convincing Evidence* is provided for Grades 4 through 6 but not Grade 3.

Summary of Criterion-Related Evidence for Validity

First, additional predictive-related validity evidence for Kindergarten through Grade 2 may need to be gathered with a different criterion assessment. Since TEMA-3 is the only criterion assessment used at those grade levels, it is unclear whether the low correlations are due to poor predictive power or due to factors such as differences in tested content.

Second, similar to the results from the concurrent-related evidence for validity, the results from SAT10 and STAAR mostly align, wherein there is mostly *Partially Convincing* to *Convincing Evidence* for concurrent-related evidence for validity for Grades 3 through 6 but not for Grades 7 and 8. These grade levels warrant additional investigation, which may include a content alignment study, an investigation of the blueprint and items, or an additional criterion assessment to confirm the results.

Reliability and Validity Evidence Disaggregated by Relevant Subgroup

Reliability Disaggregated by Relevant Subgroup

Reliability for three relevant subgroups were investigated: gender, race, and economic disadvantage. Table 15 shows the sample sizes of students who were administered the BOY, MOY, and EOY ISIP Math in each grade for each of these subgroups.

Table 15
Sample sizes taking BOY, MOY, and EOY ISIP Math for relevant subgroups by grade

		Gender		Race			Economically Disadvantaged	
Administration	Grade	Female	Male	Black	Hispanic	White	No	Yes
BOY	K	83	86	16	67	79	87	82
	1	101	115	17	67	117	109	107
	2	105	112	21	70	109	102	115
	3	101	107	19	69	105	102	106
	4	70	63	8	41	79	79	54
	5	119	129	31	83	123	138	110
	6	131	116	27	95	100	143	104
	7	149	118	27	86	132	188	79
	8	99	86	27	59	84	106	79
MOY	K	78	87	16	66	77	83	82
	1	101	128	17	77	120	110	119
	2	105	106	21	69	104	100	111
	3	99	105	19	69	101	101	103
	4	71	62	8	43	77	78	55
	5	120	130	32	84	123	137	113
	6	132	126	33	97	101	150	108
	7	138	111	24	75	127	176	73
	8	111	82	27	65	85	105	88
EOY	K	76	83	16	61	77	81	78
	1	98	122	16	75	114	110	110
	2	102	105	21	68	102	98	109
	3	96	102	18	65	100	99	99
	4	72	60	7	42	78	79	53
	5	121	130	32	85	123	136	115
	6	127	122	29	93	101	143	106
	7	135	102	22	77	115	165	72
	8	111	88	27	67	86	119	80

When interpreting the results of disaggregated coefficients of reliability and validity, it is to be noted that the sample sizes in some of the subgroups, such as the Black/African American ethnicity/race, are small. Limited interpretations can be made with such small sample sizes.

The reliability evidence in Table 10 was disaggregated by three relevant subgroups – gender, race, and economic disadvantage. Tables 16-18 show the BOY, MOY, and EOY ISIP Math reliability coefficients disaggregated by the relevant subgroups, respectively.

Table 16
Evidence for reliability of BOY ISIP Math disaggregated by relevant subgroup

		Gender		Race			Economically Disadvantaged	
Grade	Overall	Female	Male	Black	Hispanic	White	No	Yes
K	.89	.86	.91	.91	.88	.90	.89	.89
1	.92	.92	.92	.94	.93	.91	.90	.93
2	.90	.90	.90	.86	.91	.90	.89	.91
3	.90	.91	.89	.83	.92	.90	.89	.91
4	.85	.83	.88	.89	.78	.85	.84	.85
5	.92	.91	.92	.91	.91	.91	.91	.91
6	.81	.80	.81	.85	.82	.76	.77	.83
7	.85	.84	.85	.89	.88	.72	.83	.85
8	.83	.82	.84	.84	.82	.82	.85	.80

Overall, there appear to be subtle differences in reliabilities across and within the different subgroups taking the BOY ISIP Math. The data indicate that the measures of scaled scores on the BOY ISIP Math are reliable estimates and are comparable across subgroups. Using the NCRTI (2010) criterion, Grade 4 Hispanic students, Grade 6 and Grade 7 White students, and Grade 6 non-economically disadvantaged students do not provide acceptable reliability coefficients.

Table 17
Evidence for reliability of MOY ISIP Math disaggregated by relevant subgroup

		Gender		Race			Economically Disadvantaged	
Grade	Overall	Female	Male	Black	Hispanic	White	No	Yes
K	.92	.90	.94	.92	.92	.92	.91	.93
1	.94	.94	.94	.92	.95	.94	.94	.94
2	.92	.92	.92	.94	.91	.92	.92	.92
3	.88	.87	.88	.88	.90	.86	.85	.89
4	.88	.85	.90	.91	.85	.86	.86	.88
5	.87	.87	.88	.87	.88	.84	.82	.90
6	.85	.83	.87	.86	.84	.80	.83	.87
7	.88	.88	.87	.93	.90	.81	.86	.89
8	.88	.88	.89	.85	.89	.88	.89	.86

The data in Table 17 indicate that the measures of scaled scores on the MOY ISIP Math are reliable estimates and are comparable across subgroups. Overall, there only appear to be subtle differences in reliabilities across and within the different subgroups taking the MOY ISIP Math. Using the NCRTI (2010) criterion, all subgroups provide *Partially Convincing Evidence* of reliability.

Table 18
Evidence for reliability of EOY ISIP Math disaggregated by relevant subgroup

		Gender		Race			Economically Disadvantaged	
Grade	Overall	Female	Male	Black	Hispanic	White	No	Yes
K	.95	.93	.96	.94	.95	.94	.95	.95
1	.94	.95	.94	.95	.95	.93	.93	.95
2	.93	.93	.93	.91	.93	.94	.93	.93
3	.94	.94	.94	.95	.95	.94	.94	.95
4	.93	.92	.95	.93	.91	.93	.91	.95
5	.92	.92	.91	.88	.92	.90	.88	.93
6	.92	.91	.92	.91	.91	.90	.92	.90
7	.93	.92	.94	.93	.94	.88	.91	.93
8	.91	.90	.93	.92	.91	.92	.92	.90

Similar to the MOY disaggregation of reliabilities, the data in Table 18 indicate that the measures of scaled scores on the EOY ISIP Math are reliable estimates and are comparable across subgroups. Using the NCRTI (2010) criterion, all subgroups provide *Partially Convincing Evidence* of reliability.

Concurrent-Related Evidence for Validity Disaggregated by Relevant Subgroup

The evidence for concurrent-related validity presented in Table 11 is also disaggregated by the three relevant subgroup in Table 19.

Table 19
Concurrent-related evidence for validity disaggregated by relevant subgroup

		Overall	Gender		Race			Economically Disadvantaged	
Assessment	Grade	Coefficient	Male	Female	Black	Hispanic	White	No	Yes
STAR Math (BOY)	1	.66	.66	.67	.23	.70	.71	.69	.63
	2	.76	.77	.73	.82	.76	.74	.81	.72
	3	.71	.77	.63	.61	.80	.67	.64	.76
	4	.64	.57	.70	-.57	.78	.63	.51	.77
	5	.55	.55	.54	.36	.49	.57	.55	.50
	6	.74	.80	.65	.82	.74	.65	.70	.76
	7	.61	.65	.57	.26	.52	.66	.60	.55
	8	.61	.60	.62	.70	.55	.55	.63	.57
STAR Math (MOY)	1	.77	.77	.78	.77	.78	.77	.79	.74
	2	.81	.81	.81	.85	.79	.84	.82	.80
	3	.75	.78	.69	.93	.78	.71	.73	.76
	4	.67	.68	.69	-.10	.73	.67	.64	.75
	5	.71	.71	.71	.81	.63	.68	.64	.75
	6	.77	.78	.75	.70	.77	.70	.80	.70
	7	.60	.55	.64	.81	.66	.35	.46	.77
	8	.59	.60	.59	.47	.64	.59	.61	.59
STAR Math (EOY)	1	.72	.69	.75	.80	.66	.73	.71	.71
	2	.75	.80	.71	.63	.80	.78	.76	.76
	3	.74	.80	.65	.71	.81	.70	.76	.72
	4	.78	.79	.77	1.00	.81	.78	.67	.86
	5	.76	.76	.77	.68	.79	.71	.64	.80

	6	.80	.80	.80	.67	.83	.76	.76	.84
	7	.76	.78	.74	.84	.74	.58	.71	.78
	8	.61	.71	.53	.55	.71	.60	.62	.60
	K	.49	.55	.43	.17	.43	.66	.45	.44
TEMA-3	1	.66	.66	.67	.62	.64	.68	.64	.64
	2	.69	.72	.65	.75	.68	.69	.74	.62
	3	.82	.86	.77	.83	.87	.79	.82	.82
	4	.82	.82	.82	.72	.77	.86	.77	.83
SAT10	5	.82	.83	.81	.72	.79	.81	.78	.81
	6	.83	.81	.86	.79	.78	.82	.84	.80
	7	.57	.64	.50	.69	.61	.44	.50	.67
	8	.67	.72	.63	.73	.56	.72	.73	.54
	3	.82	.84	.79	.86	.84	.81	.83	.82
	4	.82	.78	.84	.76	.75	.85	.77	.82
SAT10 PS	5	.75	.75	.74	.59	.71	.73	.69	.74
	6	.83	.82	.85	.79	.79	.83	.85	.80
	7	.45	.51	.40	.58	.59	.23	.33	.63
	8	.65	.73	.56	.68	.53	.69	.75	.46
	3	.69	.75	.63	.65	.78	.66	.65	.72
	4	.71	.74	.67	.64	.70	.73	.65	.73
SAT10 P	5	.78	.82	.78	.76	.79	.73	.73	.79
	6	.74	.71	.78	.58	.67	.74	.75	.70
	7	.58	.64	.51	.73	.49	.58	.56	.60
	8	.54	.58	.51	.59	.48	.59	.53	.52
	3	.81	.82	.79	.87	.87	.75	.79	.81
	4	.80	.81	.80	.90	.83	.73	.68	.87
STAAR	5	.81	.77	.84	.74	.80	.78	.74	.83
	6	.85	.85	.85	.82	.81	.85	.85	.83
	7	.70	.70	.71	.68	.78	.54	.62	.74
	8	.68	.61	.73	.39	.73	.68	.69	.64

In general, concurrent-related coefficients for validity are stable across subgroups with similar sizes and are less than .15 from each other. Exceptions to this include differences in coefficients for economic disadvantage seen with Grade 4 (BOY STAR Math), Grade 7 (MOY STAR Math), and Grades 4 and 5 (EOY STAR Math). There are also differences in the coefficients for gender and economic disadvantage for Grades 7 and 8 on SAT10. Given that there is evidence for the technical adequacy of Grades 4 and 5 using other sources, low disaggregated correlations may possibly be due to the items on the criterion itself. However, low disaggregated correlations on Grade 7 and Grade 8 may need further investigation for gender-biased and economically biased items (i.e., items that focus on money and/or financial literacy).

The NCRTI (2010) established a criterion of .70 by *any* subgroup for demonstrating acceptable concurrent associations with the criterion assessment. Due to the sparse sample sizes noted in Table 15 for race and because economic disadvantage was a proxy variable, gender was used to determine the level of concurrent-related evidence for validity disaggregated by demographic subgroup. Table 20 summarizes the level of evidence based on the NCRTI (2010) criteria.

Table 20

Summary of concurrent-related evidence for validity disaggregated by demographic subgroup (gender) based on the NCRTI (2010) criteria

	TEMA-3	SAT10	SAT10 P	SAT10 PS	STAAR	STAR Math
K	Unconvincing					
1	Unconvincing					Unconvincing
2	Unconvincing					Partially Convincing
3		Partially Convincing	Unconvincing	Partially Convincing	Partially Convincing	Partially Convincing
4		Convincing	Unconvincing	Partially Convincing	Convincing	Unconvincing
5		Convincing	Partially Convincing	Unconvincing	Partially Convincing	Unconvincing
6		Convincing	Unconvincing	Convincing	Convincing	Partially Convincing
7		Unconvincing	Unconvincing	Unconvincing	Unconvincing	Unconvincing
8		Unconvincing	Unconvincing	Unconvincing	Unconvincing	Unconvincing

Before disaggregation, many coefficients provided *Partially Convincing Evidence* to *Convincing Evidence* (see Table 12), disaggregating the coefficients changed the level of evidence provided. Investigating the items for interactions with subgroups may provide insights into these observed results.

Predictive-Related Evidence for Validity Disaggregated by Relevant Subgroup

The evidence for predictive-related validity presented in Table 13 is also disaggregated by the three relevant subgroup and presented in Tables 21 and 22 (for the BOY and MOY ISIP Math coefficients, respectively).

Table 21

Predictive-related evidence for validity disaggregated by relevant subgroup (BOY ISIP Math)

			Gender		Race			Economically Disadvantaged	
Assessment	Grade	Overall Coefficient	Male	Female	Black	Hispanic	White	No	Yes
TEMA-3	K	.51	.60	.39	.30	.41	.65	.52	.48
	1	.65	.61	.71	.37	.72	.65	.65	.63
	2	.60	.65	.56	.65	.68	.53	.52	.66
SAT10	3	.73	.76	.71	.81	.79	.73	.73	.73
	4	.71	.66	.77	.78	.62	.74	.67	.73
	5	.57	.62	.52	.33	.61	.51	.57	.53
	6	.72	.73	.70	.64	.66	.75	.72	.70
	7	.43	.45	.43	-.06	.51	.46	.43	.45
	8	.62	.56	.66	.77	.56	.53	.56	.68
SAT10 PS	3	.75	.78	.71	.77	.78	.75	.72	.77
	4	.72	.67	.77	.81	.63	.75	.69	.72
	5	.58	.62	.51	.30	.58	.54	.59	.51
	6	.72	.72	.73	.67	.65	.76	.71	.72
	7	.31	.34	.31	-.10	.45	.31	.25	.41

	8	.75	.56	.68	.81	.50	.52	.58	.68
SAT10 P	3	.60	.65	.58	.69	.70	.60	.59	.62
	4	.58	.54	.64	.71	.51	.60	.53	.61
	5	.48	.56	.45	.30	.57	.39	.45	.46
	6	.63	.66	.60	.41	.57	.69	.66	.58
	7	.48	.48	.48	.02	.49	.50	.53	.43
	8	.50	.47	.52	.67	.52	.42	.43	.58
STAAR	3	.65	.62	.68	.54	.74	.66	.62	.68
	4	.70	.67	.76	.90	.66	.66	.58	.82
	5	.54	.53	.55	.51	.52	.47	.53	.51
	6	.72	.74	.70	.81	.69	.72	.73	.72
	7	.69	.66	.70	.67	.68	.61	.64	.72
	8	.57	.55	.58	.46	.47	.61	.64	.40

In general, predictive-related coefficients for validity (BOY) are stable across subgroups with similar sizes and are less than .15 from each other. Exceptions to this are differences between the coefficients for gender seen in Kindergarten (TEMA-3), economic disadvantage in Grades 7 and 8 on SAT10, and economic disadvantage in Grade 4 and 8 on the STAAR. Since TEMA-3 is the only criterion assessment that provides predictive-related evidence for Grade 2, items on BOY Grade 2 ISIP Math may need to be further investigated. As stated earlier, Grades 7 and 8 need further investigation.

Table 22
Predictive-related evidence for validity disaggregated by relevant subgroup (MOY ISIP Math)

								Economically Disadvantaged	
		Overall	Gender		Race				
Assessment	Grade	Coefficient	Male	Female	Black	Hispanic	White	No	Yes
TEMA-3	K	.44	.51	.39	.23	.34	.62	.43	.46
	1	.74	.75	.72	.70	.73	.73	.74	.71
	2	.70	.71	.70	.74	.71	.69	.66	.74
SAT10	3	.71	.75	.68	.71	.77	.69	.71	.71
	4	.76	.73	.77	.64	.71	.78	.71	.77
	5	.75	.74	.77	.71	.77	.69	.68	.77
	6	.72	.73	.72	.64	.61	.76	.74	.69
	7	.51	.55	.49	.68	.58	.29	.41	.60
	8	.67	.71	.60	.77	.54	.72	.67	.62
SAT10 PS	3	.73	.75	.70	.75	.77	.71	.70	.75
	4	.75	.71	.78	.66	.71	.75	.70	.77
	5	.72	.70	.75	.69	.72	.65	.66	.73
	6	.73	.74	.72	.64	.64	.77	.75	.69
	7	.43	.46	.42	.65	.62	.11	.30	.57
	8	.64	.70	.55	.74	.50	.67	.68	.54
SAT10 P	3	.58	.64	.56	.51	.65	.58	.58	.59
	4	.65	.63	.66	.56	.63	.67	.60	.66
	5	.67	.69	.68	.61	.74	.58	.58	.70
	6	.63	.62	.65	.47	.44	.69	.66	.58
	7	.48	.51	.47	.62	.42	.43	.44	.52
	8	.54	.56	.50	.66	.46	.56	.49	.57

STAAR	3	.68	.64	.73	.68	.77	.63	.65	.71
	4	.77	.79	.78	.84	.79	.72	.68	.83
	5	.74	.74	.75	.66	.76	.69	.65	.79
	6	.79	.81	.77	.80	.74	.75	.78	.81
	7	.66	.68	.67	.88	.74	.55	.61	.73
	8	.65	.61	.68	.36	.61	.66	.67	.59

In general, predictive-related coefficients for validity (MOY) are stable across subgroups with similar sizes and are less than .15 from each other. Exceptions to this are differences between the coefficients for gender seen in Grade 8 on SAT10 and economic disadvantage in Grades 7 and 8 on SAT10.

The NCRTI (2010) established a criterion of .70 by *any* subgroup for demonstrating acceptable predictive associations with the criterion assessment. Due to the sparse sample sizes noted in Table 15 for race and because economic disadvantage was a proxy variable, gender was used to determine the level of predictive-related evidence for validity disaggregated by demographic subgroup (similar to correlations with concurrent-related evidence for validity disaggregated by demographic subgroup).

Table 23 summarizes the level of evidence across Tables 21 and 22 based on the NCRTI (2010) criteria.

Table 23
Summary of predictive-related evidence for validity disaggregated by demographic subgroup (gender) based on the NCRTI (2010) criteria

	TEMA-3	SAT10	SAT10 P	SAT10 PS	STAAR
K	Unconvincing				
1	Partially Convincing				
2	Partially Convincing				
3		Partially Convincing	Unconvincing	Convincing	Unconvincing
4		Partially Convincing	Unconvincing	Partially Convincing	Partially Convincing
5		Partially Convincing	Unconvincing	Partially Convincing	Partially Convincing
6		Convincing	Unconvincing	Convincing	Convincing
7		Unconvincing	Unconvincing	Unconvincing	Unconvincing
8		Unconvincing	Unconvincing	Unconvincing	Unconvincing

Conclusions

This study collected evidence to evaluate the appropriateness of Istation's Indicators of Progress (ISIP) Math for making screening decisions for students in Kindergarten through Grade 8. Evidence was evaluated according to the criteria proposed by the NCRTI (2010). Evidence gathered includes (a) generalizability of the sample, (b) classification accuracy of the performance level, (c) evidence for reliability, (d) evidence for validity, and (e) evidence for reliability and validity disaggregated by relevant subgroup.

The generalizability of the sample is moderate as indicated in Table 3. The sample is very similar to both the statewide and national proportions for gender and economically disadvantaged status. While the sample population is comparable to the state and national population with regard to most racial/ethnic groups, the sample has a lower Hispanic and a higher White composition than the state; it also has a higher Hispanic and a lower White composition compared to the nation. Therefore, the results of this study may be generalizable to the larger student population of Texas and across the nation.

As summarized in Table 9, the classification accuracy of ISIP Math is stronger from Grades 3 through 6 than at Grades 7 and 8 based on the criterion measures selected. While the SAT10 Procedures subtest provides *Unconvincing Evidence* at all grade levels, this could be attributed to assessing different cognitive engagement levels. For example, ISIP Math items may assess multiple cognitive demands, while the SAT10 Mathematics Procedures subtest may include only procedural items. The evidence for classification accuracy of ISIP Math using the TEMA-3 was deemed *Unconvincing* to *Partially Convincing*. Without using another criterion measure, it is difficult to understand whether issues exist with the items on K-2 ISIP Math, or whether the criterion chosen was not aligned with ISIP Math (i.e., content, cut score, etc.). Across all grade levels, it was noted that ISIP Math classifies more students as "at-risk" than the criterion assessments. While over-classification may be problematic from a systems-perspective (e.g., resource allocation), it is less problematic for a screener to over-identify students as "at-risk" than it is to misclassify at-risk students who needed additional support as "not-at-risk." Also, as stated earlier, all interpretations made from this study should note that while ISIP Math uses a standard criterion for differentiating scores across tiers, different summative assessments have different cut scores, which may have contributed to varying levels of evidence.

As shown in Table 10, across all grade levels and administrations, reliability meets the .80 threshold set by NCRTI (2010). However, there is only *Partially Convincing Evidence* for reliability of ISIP Math based on their criteria given that only one reliability coefficient was presented.

As reflected in Table 12, additional concurrent-related validity evidence for K-2 may need to be gathered with a different criterion assessment since the results from TEMA-3 and STAR Math are contradictory. TEMA-3 provides *Unconvincing Evidence* at all grade levels, while STAR Math provides *Partially Convincing* to *Convincing Evidence* at Grades 1 and 2. For Grades 3-8, the correlations between ISIP Math and SAT10 and STAAR mostly moderate to strong. There is *Convincing Evidence* for concurrent-related evidence for validity from Grades 3 through 6 but not necessarily for Grades 7 and 8. These grade levels may need a deeper analysis of the content,

including an investigation of the blueprint and items, or may require an additional criterion assessment to confirm the results.

As shown in Table 14, predictive-related validity evidence for K-2 may also need to be gathered with a different criterion assessment since the evidence from the only criterion assessment used at these grades, the TEMA-3, provided *Unconvincing* to *Partially Convincing* evidence. Similar results were observed for Grades 3-8. Moderate to strong correlations between ISIP Math and SAT10 and STAAR were obtained. The predictive-related validity evidence provides *Partially Convincing* to *Convincing Evidence* for Grades 3 through 6 but not for Grades 7 and 8. These grade levels may need a deeper analysis of the content, a content alignment study, an investigation of the blueprint and items, or an additional criterion assessment to confirm the results.

As demonstrated in Tables 16-18, while there are subtle differences in reliability coefficients across and within the different subgroup in different administrations of ISIP Math, the data indicate that the measures of scaled scores on the ISIP Math are reliable estimates and are comparable across subgroups, providing *Partially Convincing Evidence* (because of only one estimate of reliability above .80). A few disaggregated coefficients for race and gender at the BOY administration are below this threshold.

As summarized in Tables 20 and 23, the evidence for validity disaggregated by subgroup follows similar trends as the evidence presented in Table 14. Grades 7 and 8 provide *Unconvincing Evidence* for validity disaggregated by subgroup across all criterion measures. Kindergarten through Grade 2 needs further investigation since only one criterion measure was used for predictive-related validity evidence, and the evidence for concurrent-related validity is inconclusive.

Overall, the evidence suggests the following:

- The generalizability and reliability of ISIP Math within this study is moderate to strong across all grade levels. Presenting another coefficient of reliability would be recommended and may improve the level of evidence from *Partially Convincing Evidence* to *Convincing Evidence*. Additionally, collecting data from a sample that has a closer representation of White and Hispanic students compared to the state and nation would improve the generalizability of the results.

- More evidence needs to be gathered for the technical adequacy of the Kindergarten ISIP Math using another criterion assessment and larger sample sizes. Doing so would provide additional evidence beyond the results obtained from the TEMA-3.

- For Grades 1 and 2, there is conflicting evidence presented for the concurrent-related evidence for ISIP Math. TEMA-3 provides *Unconvincing Evidence* at all grade levels, and STAR Math provides *Partially Convincing* to *Convincing Evidence*. More evidence needs to be gathered for Grades 1 and 2 to strengthen the technical adequacy of Grades 1 and 2 ISIP Math for all administrations (BOY, MOY, and EOY).

- For Grades 3 through 6, there is *Partially Convincing* to *Convincing Evidence* of appropriate classification of "at-risk" or "not-at-risk" by ISIP Math using SAT10, the SAT10 Mathematics Problem Solving subtest, and the STAAR. There is also evidence for concurrent- and predictive-related validity at these grade levels. While convincing evidence is not provided by the SAT10 Mathematics Procedures subtest, this may be due to differences in content between the assessments. For example, while the SAT10 Mathematics Procedures subtest may emphasize procedures, ISIP Math assess four cognitive engagement levels, including procedural fluency. Also, STAR Math does not provide convincing validity evidence at Grade 3, which may be due to lack of content alignment between ISIP Math and STAR Math at this grade level.

- Across all administrations, Grades 7 and 8 provide *Unconvincing Evidence* of classification accuracy, concurrent-related validity evidence, and predictive-related validity evidence. Coefficients disaggregated by relevant subgroup are also unstable in many cases. One possible reason may be the change in standards in 2013-14, which may have impacted student scores in having them have less time in preparing for content that aligns with these standards. This needs to be further investigated.

References

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

Bechger, T. M., Maris, G., Verstralen, H. H. F. M., & Beguin, A. A. (2003). Using Classical Test Theory in combination with Item Response Theory. *Applied Psychological Measurement, 27*(5), 319-334.

Ginsburg, H. P., & Baroody, A. J. (2003). *Test of Early Mathematics Ability, Third Edition*. Pro-Ed: Austin, TX.

Glover, T. A., & Albers, C.A. (2007). Considerations for Evaluating Universal Screening Assessments. *Journal of School Psychology*, *46*(2), 117-135.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education*, *10*(1), 61-82.

Hatfield, C., Basaraba, D., Perry, L., Ketterlin-Geller, L. R. (2015). *Imagination Station (Istation): Universal Screener Development for Grades PK-1*. Dallas, TX: Southern Methodist University.

Hatfield, C., Ratliff, B., Axel, S., Basaraba, D., Ketterlin-Geller, L. R. (2015). *Imagination Station (Istation): Updates to Universal Screener Instrument Development for Grades 2-8*. Dallas, TX: Southern Methodist University.

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,* 174-185.

Kettler, R. J., Glover, T. A., Albers, C. A., & Feeney-Kettler, K. A. (2014). An introduction to universal screening in educational settings. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.) *Universal screening in educational settings* (pp. 3-16). Washington, DC: APA.

Kline, P. (2000). *The handbook of psychological testing*. London, UK: Routledge.

National Center on Response to Intervention. (2010). *Users guide to Universal Screening Tools Chart*. U.S. Department of Education, Office of Special Education Programs, National Center on Response to Intervention.

National Research Council (2001). *Adding it up: Helping children learn mathematics*. J. Kilpatrick, J. Swafford, & B. Findell (Eds.). Mathematics Learning Study Committee, Center for Education, Division of Behavioral and Social Sciences and Education.
Washington, DC: National Academy Press.

Pearson. (2003). *Stanford Achievement Test Series, Tenth Edition*. Pearson: Austin, TX.

R Core Team. (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Renaissance Learning. (2015). *STAR Math technical manual*. Wisconsin Rapids, WI: Author.

Shivraj, P., Bell, J., Perry, L., Zhao, M. & Ketterlin-Geller, L. R. (2016). *Imagination Station (Istation): Istation's Indicators of Progress (ISIP) Math Validity Studies*. Dallas, TX: Southern Methodist University.

Texas Education Agency. (2013). State of Texas Assessments of Academic Readiness (STAAR) Assessments: Standard Setting Technical Report. Austin, TX: Author.

Texas Education Agency. (2015). *PEIMS standard reports: Student enrollment reports*. Austin, TX: Author.

U.S. Census Bureau. (2014). *Enrollment Status of the Population 3 Years Old and Over, by Sex, Age, Race, Hispanic Origin, Foreign Born, and Foreign-Born Parentage*. Washington, DC: Author.

U.S. Department of Education, National Center for Education Statistics, & Common Core of Data. (2012). *Public elementary/secondary school universe survey*. Washington, DC: Author.

U.S. Department of Education, National Center for Education Statistics, & Common Core of Data. (2016). *National elementary and secondary enrollment by race/ethnicity projection model, 1972 through 2025*. Washington, DC: Author.