
Pathways for Instructionally Embedded Assessment

# *PIE Pilot Study: Design and Administration Evidence*

Accessible Teaching, Learning, and Assessment Systems. (2025). *PIE pilot study: Design and administration evidence.* University of Kansas; Accessible Teaching, Learning, and Assessment Systems.

# Acknowledgments

We acknowledge the contributions of many project staff members who contributed to the design and implementation of the 2024–2025 pilot study, as well as the data analysis and writing of this report. Several project members made significant writing contributions to this report: Auburn Jimenez, Jake Thompson, and Ashley Hirt. We would also like to acknowledge several project staff members for their writing contributions: Nami Shin, Allison Barkley, Eun Mi Kim, Alexus Cafferty, and Breonna Yungeberg. Breonna Yungeberg, Sarah Randol, Sarah Hardman, and Michael Muenks supported recruitment and implementation of the pilot study. Chris Lorenzen supported graphics development.

Project Director: **Shaun Bates**, Missouri Department of Elementary and Secondary Education
Principal Investigator: **Brooke Nash**, Accessible Teaching, Learning, and Assessment Systems; University of Kansas
Mathematics Pathways Director: **Mary Majerus**, Missouri Department of Elementary and Secondary Education

Project Advisory Committee (PAC)
**Brian Gong**, **Ph.D.**, *Center for Assessment*
**Bob Henson**, **Ph.D.**, *University of North Carolina at Greensboro*
**Meagan Karvonen**, **Ph.D.**, *Accessible Teaching, Learning, and Assessment Systems; University of Kansas*
**Leanne Ketterlin-Geller**, **Ph.D.**, *Southern Methodist University*
**Ed Roeber**, **Ph.D.**, *Michigan Assessment Consortium*
**Lisa Sireno**, *Missouri Department of Elementary and Secondary Education (DESE)*
**David Williamson**, **Ph.D.**, *Independent consultant*
**Phoebe Winter**, **Ph.D.**, *Independent consultant*

The project reported here is supported by the U.S. Department of Education (USED) through a Competitive Grants for State Assessments (CGSA) program under Grant No. S368A220019. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the USED and CGSA.

# Contents

# Overview

Pathways for Instructionally Embedded Assessments (PIE) is a Competitive Grants for State Assessments (CGSA)–funded project. The PIE project is led by the Missouri Department of Elementary and Secondary Education (MO-DESE) in partnership with Accessible Teaching, Learning, and Assessment Systems (ATLAS) at the University of Kansas. The project's overarching goal was to develop and evaluate a prototype instructionally embedded assessment model that is based on research-based learning pathways (LPs) and could be used for instructional and summative reporting purposes. As a proof of concept, the scope of content development for the project was focused on academic content standards in fifth grade mathematics.

The PIE project included four goals that encompass (a) design, development, and evaluation of learning pathways and aligned assessments, (b) testing usability of the system in classrooms, (c) evaluating technical adequacy of the PIE assessments including development of a theory of action and validation plan for future summative use of PIE results, and (d) dissemination of findings and products. The purpose of this report is to describe the design of the PIE assessments and pilot study administration, provide evidence of the technical adequacy of the assessments, provide administration evidence including instructionally embedded assessment use patterns, report student performance results, and summarize the evidence for each relevant claim in the theory of action.

The intended audiences for this report include state education agency staff and other researchers and assessment developers interested in through-year and other innovative assessment models.

# Introduction

The use of instructionally useful assessments to support teachers' instructional decision-making and improve student achievement is well supported by research (Choi et al., 2022; Varier et al., 2024). Teachers report that formative assessments can help them identify gaps in student learning and support their teaching but cite the sometimes-limited availability of instructionally useful assessments and the need to develop their own formative assessments for classroom use as common barriers (Martin et al., 2022). Instructionally useful assessments that are both aligned to state content standards and technically sound are often out of reach for local school districts because of resource constraints, and this lack of resources to implement such assessments can increase equity-related achievement gaps for students in low-income districts.

Current summative and interim assessment models are strictly guided by academic calendars and may not align with actual instruction and student learning. Furthermore, the type of results provided by summative assessments may not be available quickly enough for teachers to adjust instruction nor may they provide detailed information about student understanding (Marion, 2018). Misalignment of interim assessments with a curriculum can also hinder the ability of assessment data to inform instruction (Perie et al., 2007).

These limitations necessitate innovative, instructionally useful assessment systems that can combine the benefits of various models. National assessment leaders have been calling for new models of educational assessment that distribute assessments over time and make use of new scoring models while still providing instructionally useful data that can be used to improve student learning (Bennett, 2018).

When instructionally embedded assessments are used in conjunction with end-of-year testing, stakeholders are provided with additional learning opportunities throughout the school year and numerous options for tracking student progress (Speckesser et al., 2018). This approach combines multiple measures, allowing teachers to reap the benefits of traditional spring summative assessment models as well as the ongoing feedback of embedded formative assessment to better guide instruction and student learning. This model can help reduce the burden of spring testing and improve the efficiency of instruction and assessment cycles using diagnostic data (Speckesser et al., 2018).

PIE assessments measure student mastery in fifth grade mathematics. The purpose of the PIE assessments is to support classroom teachers in making instructional decisions that ultimately improve student progress toward mastery of grade-level content expectations. Results from the PIE assessments are intended to support interpretations about what students know and are able to do relative to research-based learning pathways. Results provide information that can monitor student learning and guide instructional decisions.

The scope of the PIE research and development project was fifth grade mathematics content standards. The pilot study population included fifth grade teachers who provided instruction on the Missouri priority content standards in fifth grade mathematics and their students. As part of the pilot study of the PIE system, teachers created customized content groups according to their local curriculum. For each content group, teachers administer three short assessments to their students, a baseline, a midway, and an end-of-unit assessment, during their instructional unit aligned to the content group. Teachers use assessment data during the instructional unit to pinpoint areas for additional support and adjust whole-class, small-group, or individual instruction.

## Pilot Study Goals

The overarching purpose of the pilot study was to evaluate the usability and feasibility of the PIE system in fifth grade mathematics, including teacher use of training and resources, the assessment and instruction cycles, student progress reports, and teacher perceptions of the system (see also *Educator Perspectives on the PIE Assessment System: Evidence From the PIE Pilot Study* [ATLAS, 2025b]). The data collected from the pilot study were also intended to be used to evaluate the technical adequacy of the assessments. To guide the technical evaluation of the assessments within the context of the intended uses of the assessment results, PIE project staff started by defining a theory of action.

## PIE Theory of Action and Interpretive Argument

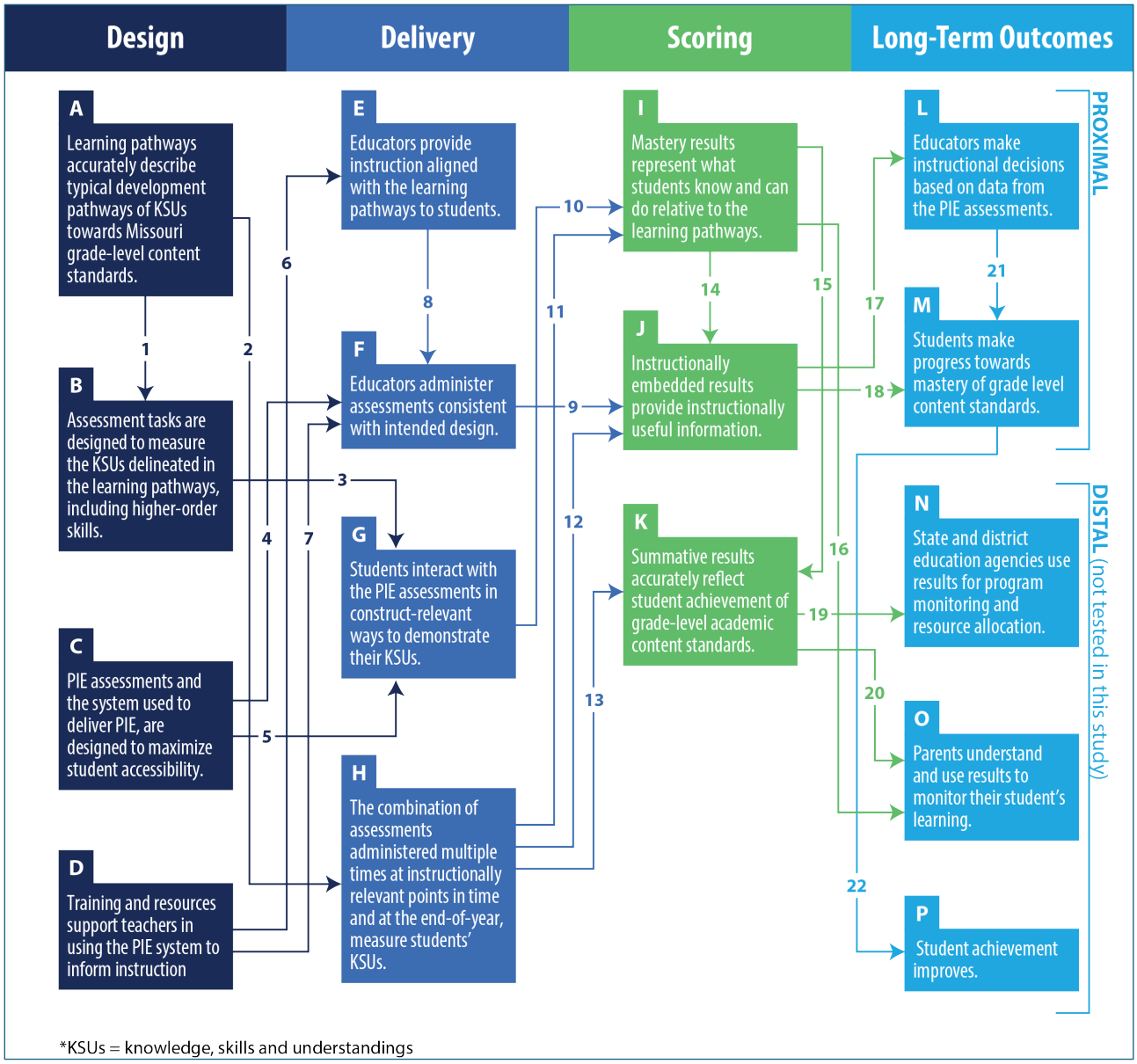
The PIE project used an argument-based approach to assessment validation which included a three-tiered approach: specification of (a) a theory of action defining claims in the validity

argument that must be in place to achieve the goals of the system, (b) an interpretive argument defining propositions that must be evaluated to support each claim in the theory of action, and (c) validity studies to evaluate each proposition in the interpretive argument (either planned as part of the current grant project or planned as a potential future study). The PIE theory of action and validation plan was used to outline the evidence needed to support valid interpretations of assessment results used for both formative and summative purposes.

The scope of evidence described in this report is limited to the design, delivery, scoring, and outcomes claims most relevant to the instructionally embedded assessments, specifically, claims D, E, F, G, H, I, and J. Additional sources of evidence for the other claims in theory of action are referenced and summarized in Chapter 6: Validity Argument of this report. Figure 1 displays the PIE Theory of Action.

Figure 1.

*PIE Theory of Action*




# Assessment Design and Development

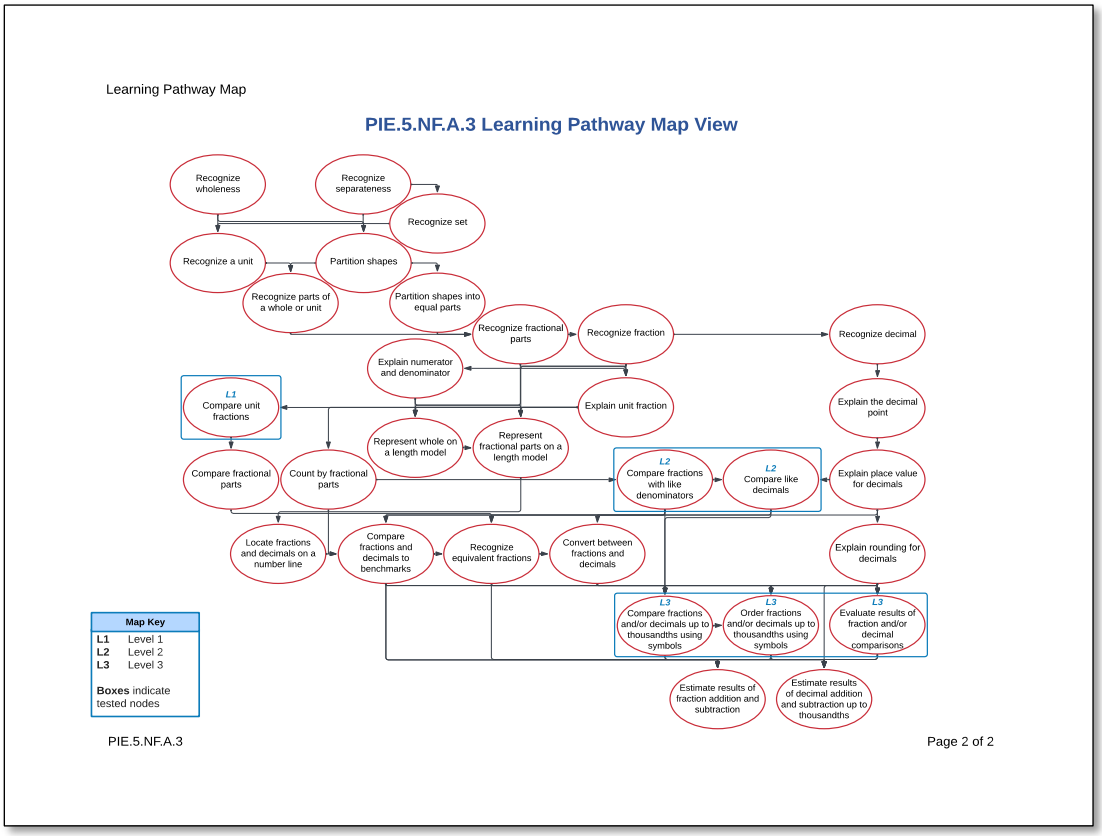
## 1.1 Assessment Structures

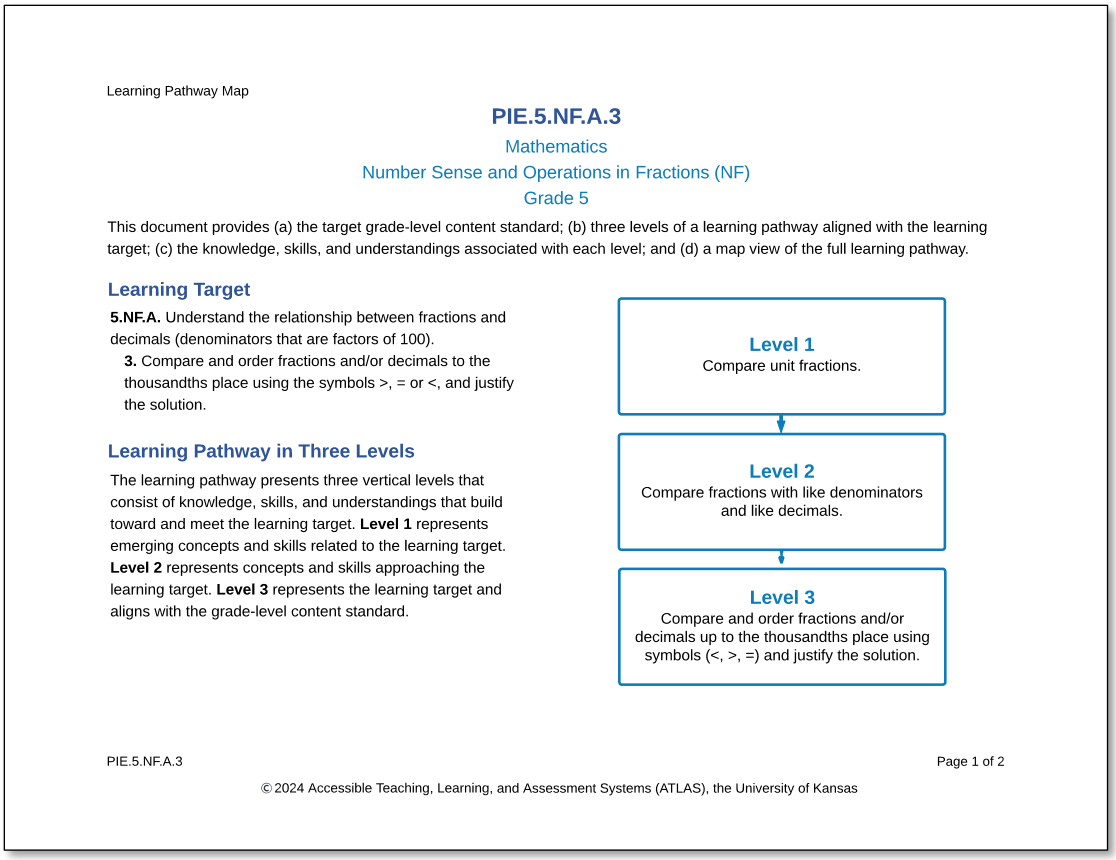
PIE learning pathways were designed and developed to serve as the foundation of the assessments and score reports. The learning pathways developed for the PIE project align to 25 fifth grade mathematics content standards. PIE staff analyzed the 25 standards and other central knowledge, skills, and understandings supporting the acquisition of the learning targets by synthesizing the available student learning and development literature understandings. The identified learning targets and supporting knowledge, skills, and understandings formed the

typical pathways of student learning toward the prioritized fifth grade mathematics content standards. The preliminary learning pathways were cross-checked with other learning maps associated with the intended learning targets. Missouri educators reviewed the content and structure of learning pathways; their feedback guided the final learning pathways that were used as the basis for subsequent assessment and reporting dashboard development. The development process resulted in 25 learning pathways (one per content standard) and 75 pathway levels (three vertical levels for each learning pathway) for use as assessment targets. Figure 2 shows an example learning pathway (map and level views) for the content standard, number sense and operations in fractions. For information about the design and development procedures of the learning pathways, refer to Kim et al. (2024).

Figure 2.

*Example PIE Learning Pathway – Map View and Level View*





Learning Pathway Map

**PIE.5.NF.A.3**

Mathematics

Number Sense and Operations in Fractions (NF)

Grade 5

This document provides (a) the target grade-level content standard; (b) three levels of a learning pathway aligned with the learning target; (c) the knowledge, skills, and understandings associated with each level; and (d) a map view of the full learning pathway.

**Learning Target**

**5.NF.A.** Understand the relationship between fractions and decimals (denominators that are factors of 100).

**3.** Compare and order fractions and/or decimals to the thousandths place using the symbols >, = or <, and justify the solution.

**Learning Pathway in Three Levels**

The learning pathway presents three vertical levels that consist of knowledge, skills, and understandings that build toward and meet the learning target. **Level 1** represents emerging concepts and skills related to the learning target. **Level 2** represents concepts and skills approaching the learning target. **Level 3** represents the learning target and aligns with the grade-level content standard.



PIE.5.NF.A.3 Page 1 of 2



Assessment items were developed to measure the knowledge, skills, and understandings for each of the 75 LP levels. In September 2023, 28 Missouri educators attended an item-writing workshop. The educators received an orientation on item-writing best practices, including fairness considerations such as Universal Design principles, accommodations, and bias, and they were asked to draft 402 items. The workshop resulted in 443 assessment items, each aligned to one of the 75 pathway levels. After the workshop, the items were further developed by the project team. This included revising them for alignment, accuracy, graphics, and accessibility. After these reviews, Missouri educators completed an external review of items. The project team then reviewed the educators' recommendations. They accepted and made no revisions to 47% of the items; 53% of items were revised according to the feedback. Ultimately, no items were rejected. The revised items then underwent additional internal reviews before being provided to state project staff for final review. In total, 411 items (including 19 item twins) were field tested with 1,708 students across Missouri. Item response data were analyzed, reviewed, and used to finalize a set of high-quality items for inclusion in Year 3 pilot study test forms. For further details of the item development process, refer to the *PIE Assessment Design and Development Technical Report* (ATLAS, 2025a).

Items that were promoted after the field test were used to populate a pool of forms that could be dynamically assigned according to teacher choice and student performance during the pilot study (see Chapter 2 for more information).

## 1.2 Evidence of Item Quality

### *1.2.1 Field Testing*

In spring 2024, an item field test was conducted in fifth grade math classrooms across Missouri to evaluate new assessment items for the 2024–2025 pilot study. A total of 1,708 students participated and 411 items, including 19 item twins (content analogs), were field tested. Item quality was evaluated using overall and conditional *p*-values, as well as discrimination indices. Using preestablished flagging criteria, PIE project staff reviewed item statistics and, in coordination with Missouri DESE, recommended items for use in the 2024–2025 pilot study. Ultimately, 334 items were approved for use in the pilot study, and 141 additional item twins were developed afterward. Each item measured a single LP level. For further details of the 2025 field test, refer to the *PIE Assessment Design and Development Technical Report* (ATLAS, 2025a).

### *1.2.2 Psychometric Properties*

Students' response data from the pilot study window were used to evaluate the psychometric properties of mathematics items to determine item quality. Several sources of evidence will be used to inform determinations about item quality, including evaluations of item difficulty and evaluations of item-level bias. In addition, item twins were evaluated and compared to their content-equivalent counterparts. Effect sizes provide one source of evidence to support the functioning of item twins in the pilot study assessment.

In total, 476 items were administered across the instructionally embedded and spring assessment windows during the 2024–2025 pilot study, and the psychometric properties of these items are discussed in the next section. In addition, we evaluated 145 item twins in comparison to their content-equivalent counterparts.

#### 1.2.2.1 Evaluation of Item Difficulty

The item-difficulty index reflects the average difficulty level of an item. For 1-point items, it corresponds to the overall probability (or *p*-value) of students answering the item correctly. For 2-point items, the index represents the average student score on the item, ranging from 0 to 2. Items were flagged if they were too challenging (indicated by a *p*-value less than .25) or were too easy (indicated by a *p*-value greater than .95).

Figure 3 and Figure 4 present the overall item-difficulty indices for 1-point and 2-point items, respectively. A minimum of 20 samples was required for inclusion in the figures to reduce the risk of skewed results. In total, no items were excluded because of sample size. Four hundred fifty-eight (96.2%) items fell within the acceptable ranges of difficulty, and 18 (3.8%) items were flagged as outside the acceptable ranges of difficulty.

Figure 3.

*Item Difficulty Indices for 1-Point Math Items*

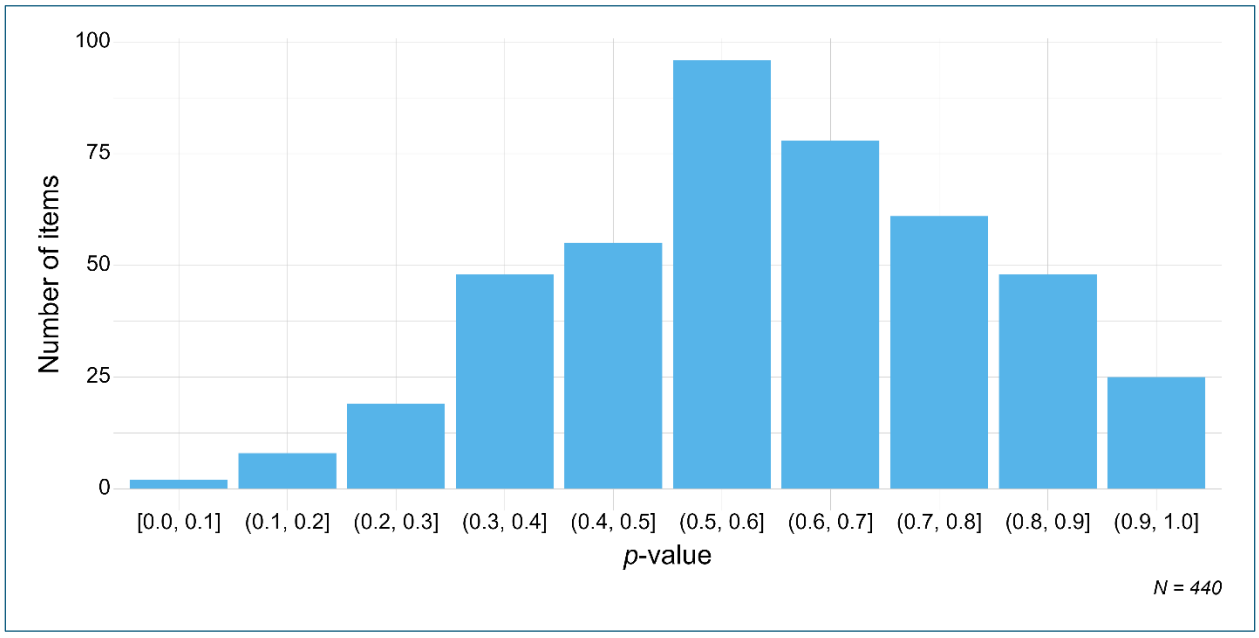
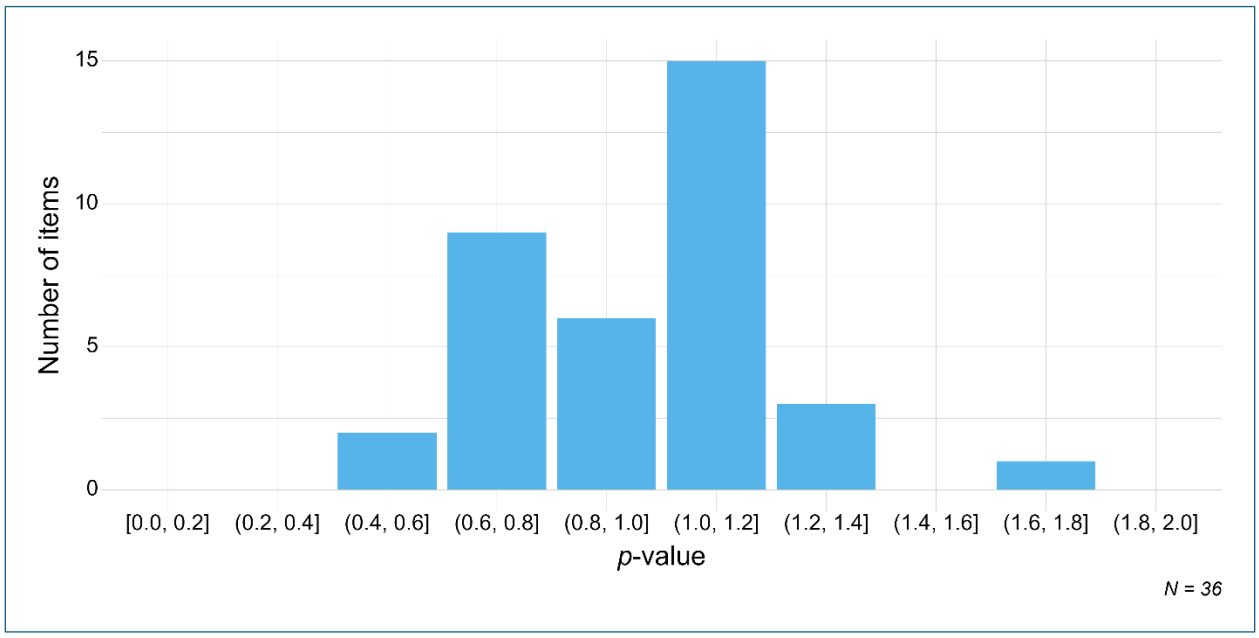



Figure 4.

*Item Difficulty Indices for 2-Point Math Items*




The *conditional item-difficulty index* (e.g., Thompson, 2019) reflects the average difficulty of an item separately for masters and nonmasters within each learning pathway level. Each item is

aligned to a specific pathway level of a content standard. For 1-point items, the index indicates the conditional probability (i.e., conditional *p*-value) that masters and nonmasters will answer the item correctly. For 2-point items, it represents the average score earned by masters and nonmasters on the item, with possible values ranging from 0 to 2. Students' response data were scored using a hierarchical diagnostic classification scoring model (HDCM; Templin & Bradshaw, 2014). The HDCM produces posterior probabilities of mastery for each student for each pathway level (discussed later in the report). A threshold of 0.7 was used to distinguish between masters and nonmasters of a pathway level. Students that exceed the threshold of 0.7 are determined as masters of a pathway level, and students that fall below the threshold of 0.7 are determined as nonmasters of a pathway level. Assessment items were flagged as potentially problematic according to two criteria. An item was flagged if masters had a conditional *p*-value below 0.4 or an average score below 0.8. Similarly, items were flagged if nonmasters had a conditional *p*-value above 0.6 or an average score exceeding 1.2.

Figure 5 and Figure 6 present the conditional item-difficulty indices for masters and nonmasters, respectively, for 1-point items. Figure 7 and Figure 8 present the conditional item-difficulty indices for masters and nonmasters, respectively, for 2-point items. A minimum of 20 samples was required for inclusion in the figures to reduce the risk of skewed results. In total, no items were excluded due to insufficient sample sizes for masters. However, 18 (3.8%) items were excluded due to insufficient sample sizes for nonmasters. 463 (97.3%) items for masters and 408 (85.7%) items for nonmasters fell within the acceptable ranges of difficulty, whereas 13 (2.7%) items for masters and 50 (10.5%) items for nonmasters were flagged as outside the acceptable ranges of difficulty.

Figure 5.

*Item Difficulty Indices for Masters for 1-Point Math Items*

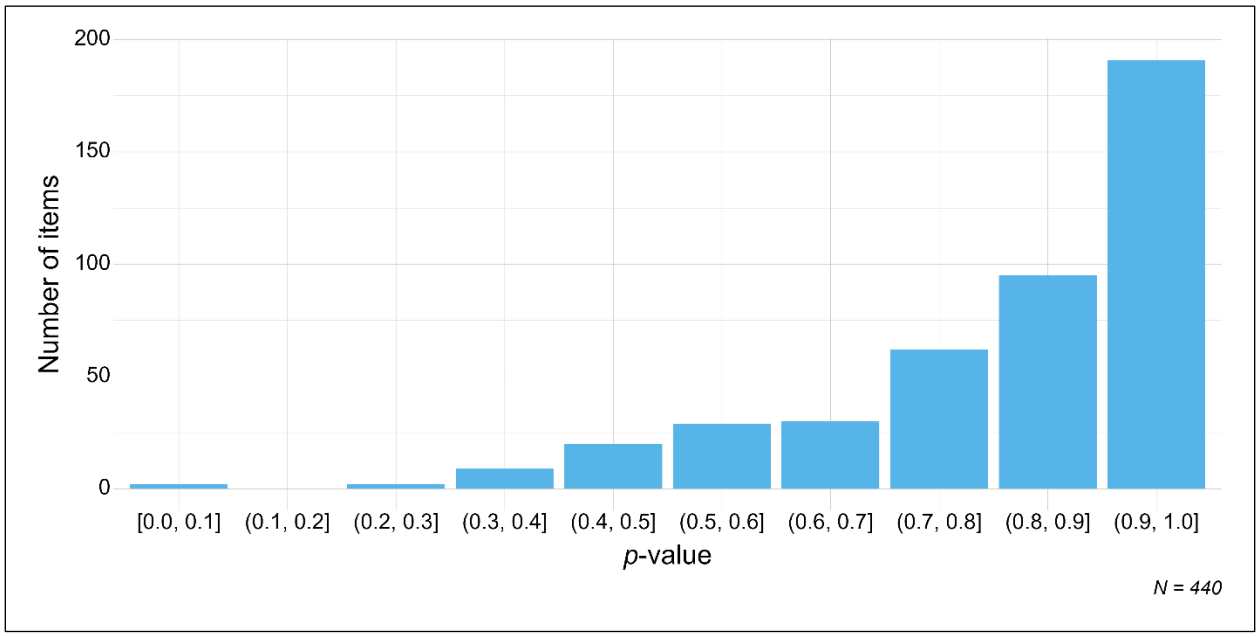


Figure 6.

*Item Difficulty Indices for Nonmasters for 1-point Math Items*

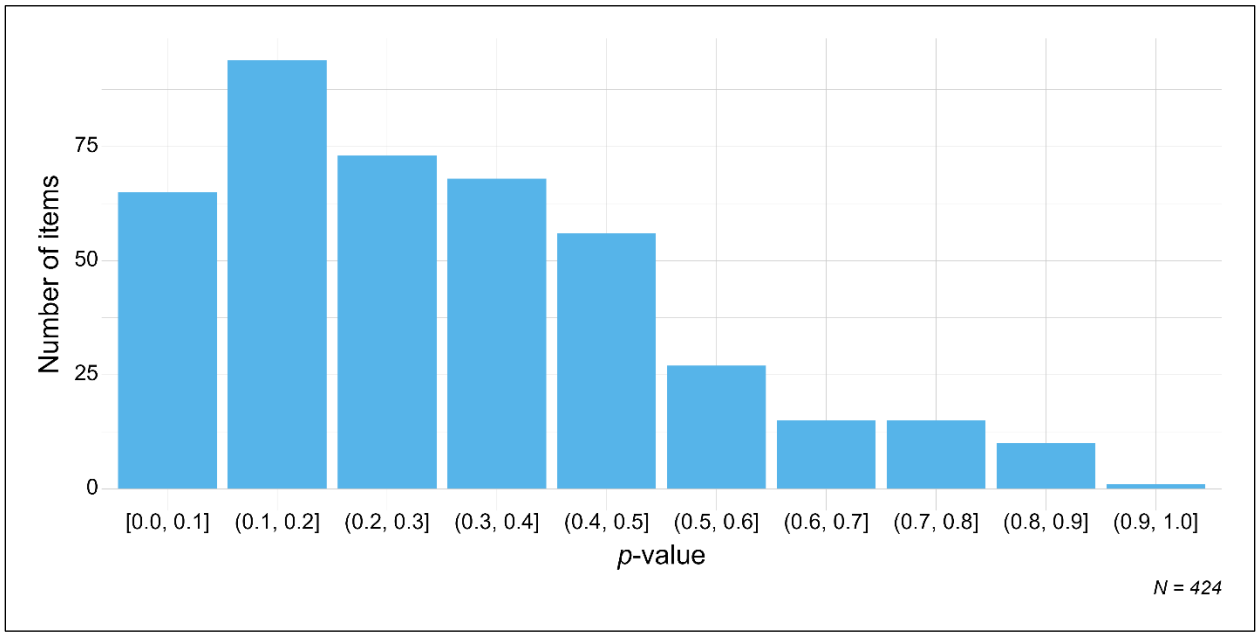



Figure 7.

*Item Difficulty Indices for Masters for 2-Point Math Items*

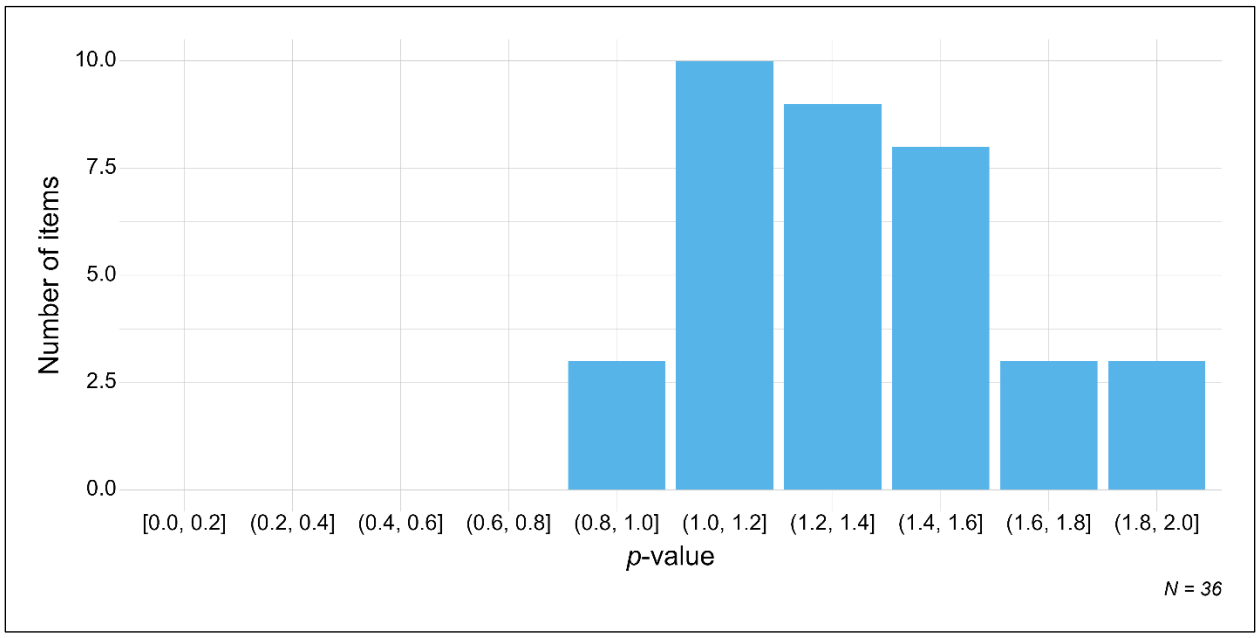
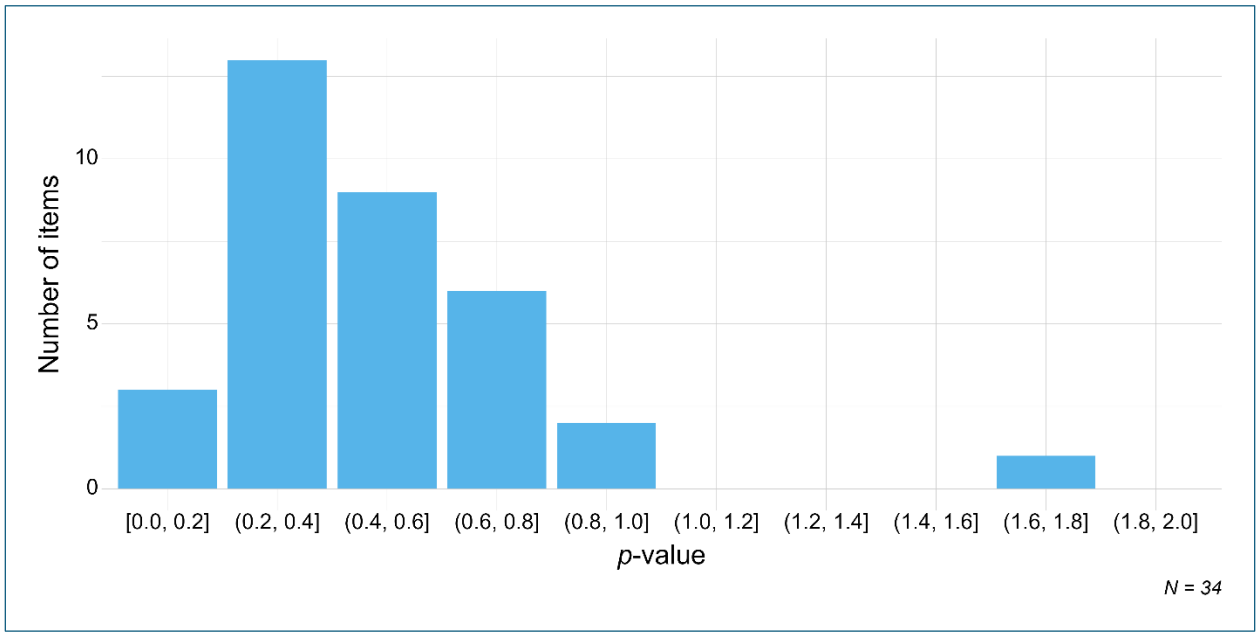


Figure 8.

*Item Difficulty Indices for Nonmasters for 2-Point Math Items*




1.2.2.2 Evaluation of Item Twins

Item twins are content-equivalent counterparts to items originally created by item writers (Karvonen et al., 2020). The item twins were developed for the PIE assessments using the task models and were used for retesting students on skills aligned to pathway levels. In total, 145 item twins were included in the PIE instructionally embedded assessments. Item twins were created using Level 1 and Level 2 items along with their corresponding task models. Each item twin targeted the same KSUs as its original counterpart and was constructed following the specifications outlined in the task model. The twins maintained the same item type and overall structure as the original items. However, certain features, such as context and specific values, were modified in the item twins.

Cohen's *h* is a standardized effect size described in Cohen (1988) that is used to quantify the magnitude of difference between two probabilities.

$$\mathrm{h} = 2\arcsin\left(\sqrt{p_1}\right) - 2\arcsin\left(\sqrt{p_2}\right). \quad (1)$$

Cohen's *h* effect size offers one method to assess whether differences between an item twin and its content-equivalent counterpart led to changes in item performance. For items scored on a 2-point scale, average scores were rescaled to fall within a 0 to 1 range, effectively converting them to probabilistic values so that the effect size could be computed using the same metric as for dichotomous items. Effect size values were reported to categorize the magnitude of differences as small, moderate, or large, using thresholds established by Cohen (1992). Specifically, values below 0.2 indicate a small effect, values between 0.2 and 0.5 represent a moderate effect, and values of 0.5 or higher reflect a large effect.

For each item, three effect size measures were calculated:

$$h_{overall} = 2\arcsin\left(\sqrt{p_{1;\,overall}}\right) - 2\arcsin\left(\sqrt{p_{2;\,overall}}\right), \quad (2)$$

$$h_{masters} = 2\arcsin\left(\sqrt{p_{1;\,masters}}\right) - 2\arcsin\left(\sqrt{p_{2;\,masters}}\right), \quad (3)$$

$$h_{nonmasters} = 2\arcsin\left(\sqrt{p_{1;\,nonmasters}}\right) - 2\arcsin\left(\sqrt{p_{2;\,nonmasters}}\right). \quad (4)$$

In equations 1-4, $p_1$ denotes the probability associated with the item twin and $p_2$ denotes the probability associated with the content-equivalent counterpart.

Figure 9 presents a breakdown of the effect size values for the overall item-difficulty indices. No effect sizes were excluded due to insufficient sample sizes. In total, 141 (97.2%) effect sizes demonstrated small or moderate effect sizes, and four (2.8%) effect sizes demonstrated large effect sizes. These findings indicate that a vast majority of the item twins functioned similarly to their content analogs.

Figure 9.

*Overall Summary of Effect Sizes*

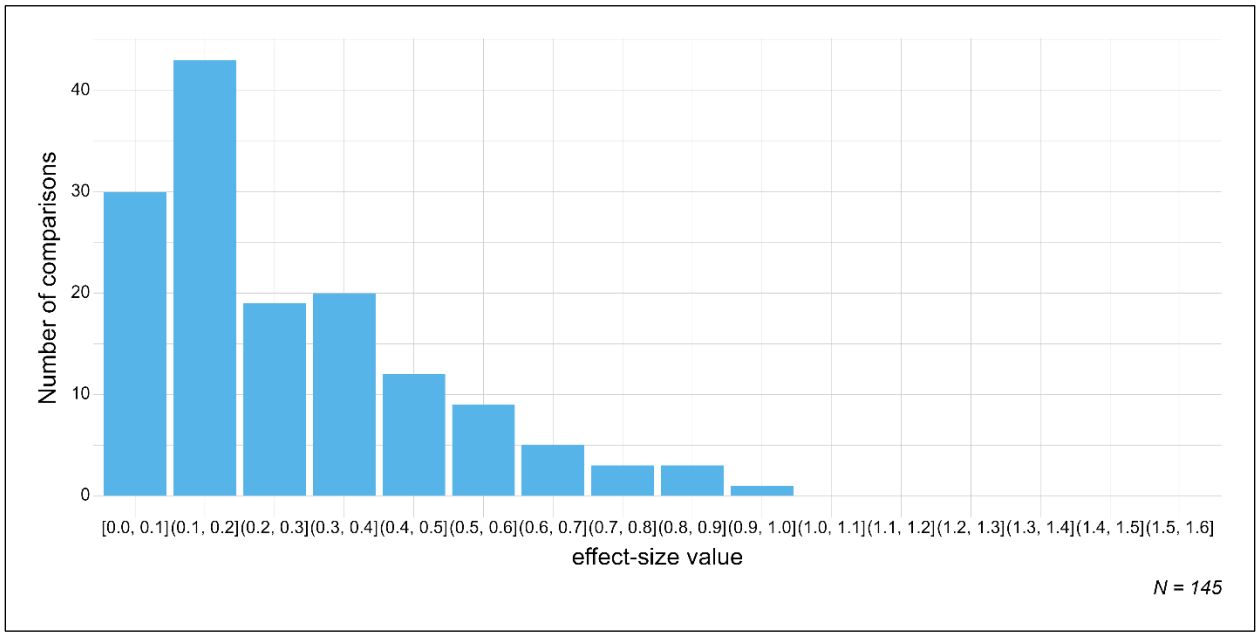



Figure 10 and Figure 11 present a breakdown of the effect size values for conditional item-difficulty indices separately for masters and nonmasters. Effect size values were included in each figure if the item twins and their content analogs met a minimum sample size of 20 students for masters and nonmasters, respectively. In total, no effect size values for masters were excluded because of insufficient sample sizes. In contrast, nine (6.2%) effect size values for nonmasters were excluded due to insufficient sample sizes. For masters, 134 (92.4%) effect sizes demonstrated small or moderate effect sizes, and 11 (7.6%) effect sizes demonstrated large

effect sizes. For nonmasters, 133 (91.7%) effect sizes demonstrated small or moderate effect sizes, and three (2.1%) effect sizes demonstrated large effect sizes.

Figure 10.

*Summary of Effect Sizes for Masters*

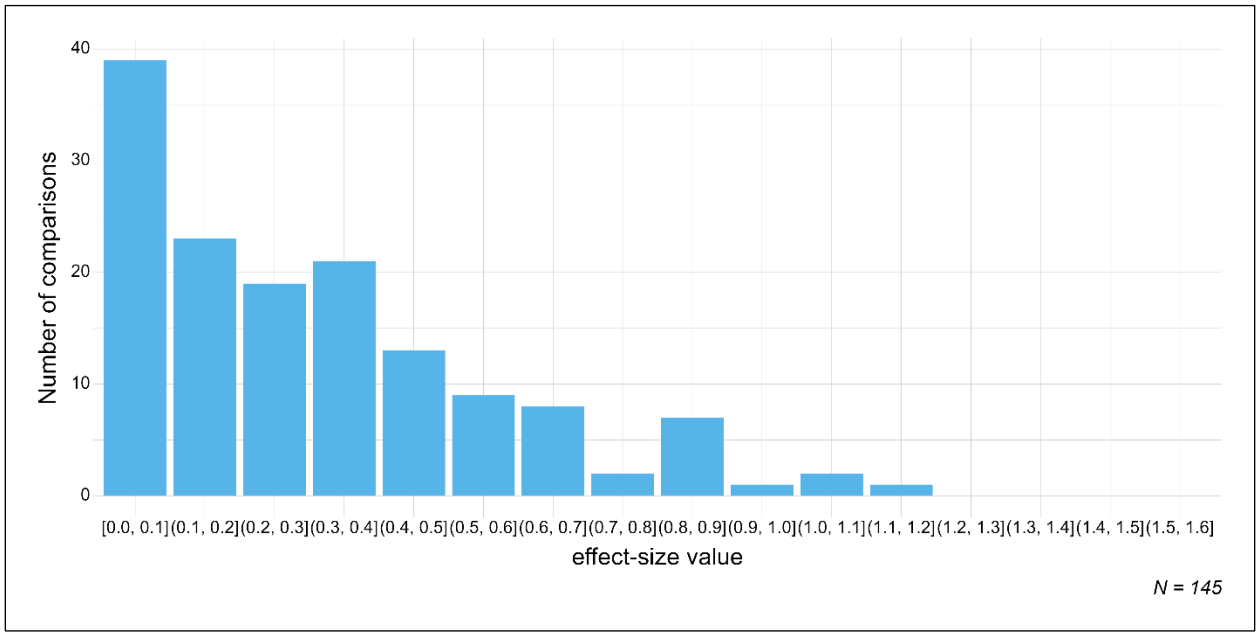
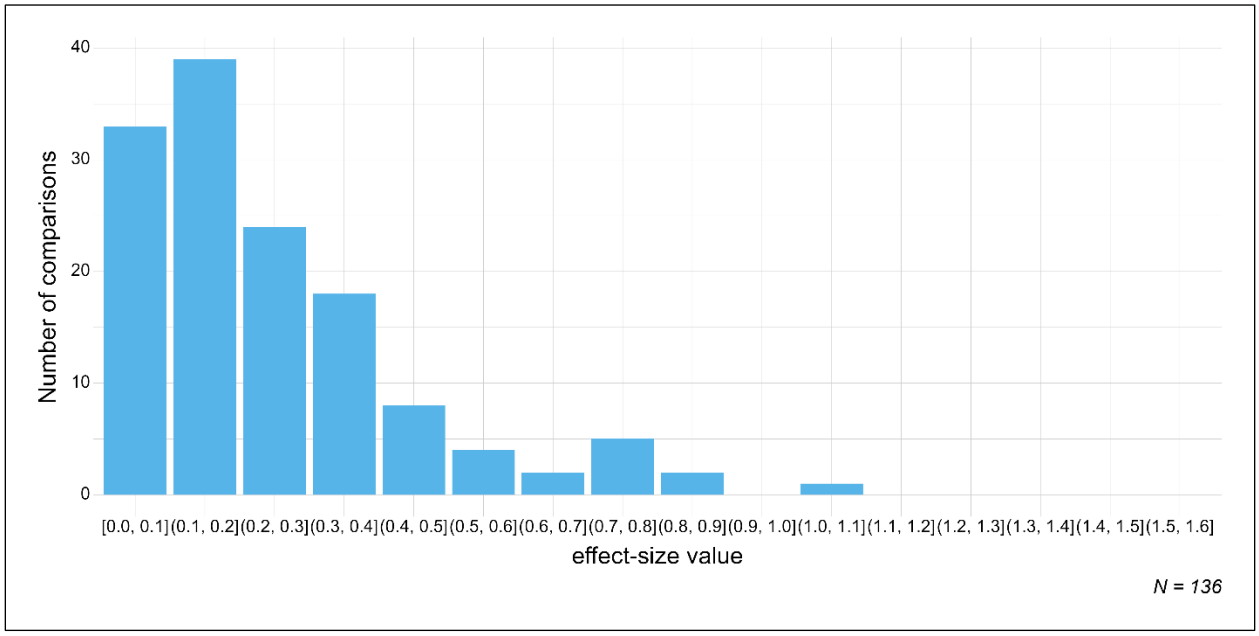



Figure 11.

*Summary of Effect Sizes for Nonmasters*



#### 1.2.2.3 Evaluation of Item-Level Bias

Another source of evidence for item quality involves evaluating potential bias in how items function across student subgroups. Differential item functioning (DIF) analyses were conducted to determine whether items performed differently according to gender and race among students who took PIE assessments during the instructionally embedded assessment window. These analyses help identify instances where items may be more difficult for certain groups, despite similar understanding of the assessed content. While the presence of DIF does not necessarily indicate a flaw in the item, it may suggest construct-irrelevant variance that could impact test fairness and validity (American Educational Research Association et al., 2014).

DIF was assessed for each item by predicting the probability of a correct response to an item given subgroup membership and a matching variable using logistic regression. DIF analyses examined how student membership in race and gender subgroups impacted item performance in the PIE assessments. Students across subgroups were matched according to their mastery status (e.g., mastery, non-mastery) on the skill measured by the item in question. In total, the DIF analyses require the estimation of three logistic regression models for each item,

$$M_0: \log i\, t(\pi_i) = \beta_0 + \beta_1 X \quad (5)$$

$$M_1: \log i\, t(\pi_i) = \beta_0 + \beta_1 X + \beta_2 G \quad (6)$$

$$M_2: \log i\, t(\pi_i) = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG \quad (7)$$

In Equations 5–7, $\pi_i$ denotes the *p*-value for item $i$ , $X$ denotes the mastery level on the assessed skill (0 = nonmaster, 1 = master), and *G* denotes the dummy-coded subgroup membership variable (0 = focal group, 1 = reference group). The item parameters in the logistic regression equations are interpreted as: $\beta_0$ denotes the intercept term, $\beta_1$ denotes the change in log-odds of a correct response associated with mastering the assessed skill, $\beta_2$ denotes the change in log-odds of a correct response for members of the reference group, and $\beta_3$ denotes the change in the log-odds of a correct response for members of the reference group who also mastered the assessed skill. When evaluating DIF across gender subgroups, the male subgroup was treated as the reference group, and the female subgroup was treated as the focal group. When evaluating DIF across race subgroups, the White subgroup was treated as the reference group, and the Black, Asian, American Indian, Native Hawaiian or Pacific Islander, Alaska Native, and two or more races subgroups were treated as the focal groups in their own respective models.

In the current study, evidence of uniform and non-uniform DIF were collected for each subgroup. Uniform DIF describes the scenario where subgroup-level item performance is consistent across mastery levels. In contrast, non-uniform DIF describes the scenario where differences in subgroup-level item performance varies by mastery level. To capture evidence of non-uniform DIF, an interaction term can be included in the logistic regression equation. The Nagelkerke pseudo $R^2$ effect size measure was calculated for each item to provide insight into the practical significance of including group membership and interaction terms into the logistic regression equations for each item. Consistent with prior operational research (Dynamic Learning Maps [DLM] Consortium, 2022), effect size thresholds described in Zumbo and Thomas (1997) and Jodoin and Gierl (2001) were used to determine whether the reported effect size

values reflected a small, moderate, or large effect. The Zumbo and Thomas (1997) thresholds consider Cohen's (1992) guidelines for identifying practical effect sizes. These thresholds define effect size values less than 0.13 as small effects, values between 0.13 and 0.26 as moderate effects, and values greater than 0.26 as large effects. The Jodoin and Gierl (2001) indices enforce stricter thresholds compared to Zumbo and Thomas (1997). These thresholds define effect size values less than 0.035 as small effects, values between 0.035 and 0.07 as moderate effects, and values greater than 0.07 as large effects.

To ensure robust DIF estimates, items were included in the analysis if they met minimum sample size thresholds and fell within an appropriate range of proportion-correct indices (*p*-values). Inclusion criteria using minimum sample size thresholds were enforced at the subgroup-level. In the PIE instructionally embedded window, the population was approximately balanced across gender subgroups (i.e., male, female), but unbalanced across race subgroups (i.e., White, Black, Asian, American Indian, Native Hawaiian or Pacific Islander, Alaska Native, and two or more races). The number of non-White students who participated in the instructionally embedded window is approximately one-third the number of White students who did; these relative proportions in sample size are expected to vary at the item level because students tested on items at different sets of items. Hence, the minimum sample size thresholds primarily affected the race subgroups that were included in the DIF analysis. Items were included in the DIF analyses when at least 50 students belonging to the focal subgroup of interest responded to the item.

Inclusion criteria using *p*-values were enforced at the overall and subgroup-levels. In the PIE instructionally embedded window, items that students overwhelmingly answered correctly or incorrectly are susceptible to estimation errors in DIF analyses, and thus were removed prior to conducting DIF analyses. Items were excluded from the DIF analyses if the overall *p*-value was greater than .95 or less than .05. Items were also excluded from the analyses if the *p*-value for one gender or race subgroup was greater than .97 or less than .03. The enforced inclusion criteria are consistent with other operational analyses conducted (e.g., DLM Consortium, 2022).

For gender, 438 (92.0%) items were included in the DIF analysis. For race, 224 (47.1%) items were included in at least one DIF analysis comparison. Participants in the pilot study represented seven race subgroups. However, only two non-White subgroups (i.e., Black, two or more races) had 50 or more students participating in the pilot study, so only two potential comparisons were possible for each item based on race subgroups. Overall, 210 (44.1%) items were evaluated for DIF for Black students, and 116 (24.4%) items were evaluated for DIF for students who identified as being of two or more races. Across all gender comparisons, sample sizes for the focal subgroup ranged from 75 to 676, and sample sizes for the reference subgroup ranged from 68 to 659. Across all race comparisons, sample sizes for the focal subgroups ranged from 50 to 108, and sample sizes for the reference subgroup ranged from 236 to 1,070.

Table 1.

*Number of Items Evaluated for Each Race Subgroup*

| Focal group | Items (*n*) |
|---|---|
| Black | 210 |
| Two or more races | 116 |

Table 2.

*Summary of Evidence of Differential Item Functioning (DIF)*

| Type of DIF | Total | $N_{Sig}$ | % | $N_{ES}$ |
|---|---|---|---|---|
| Gender | | | | |
| Uniform | 438 | 46 | 10.5 | 0 |
| Non-uniform | 438 | 44 | 10.0 | 4 |
| Race | | | | |
| Uniform | 224 | 24 | 10.7 | 2 |
| Non-uniform | 224 | 26 | 11.6 | 2 |

*Note.* Total = total number of items; $N_{Sig}$ = number of items flagged via significance testing; % = percentage of items flagged via significance testing; $N_{ES}$ = number of items with moderate or large effect size.

# Pilot Study Design

## 2.1 Population

The population of interest for the pilot study was fifth grade students and their teachers who receive and provide instruction (respectively) on the Missouri priority content standards in fifth grade mathematics.

## 2.2 Instructionally Embedded Assessment Delivery

Assessment delivery for the PIE pilot study incorporated multiple cycles of instruction and assessment designed to leverage the flexibility of the system. Educators developed custom content groups based on priority content standards and then administered a baseline (Level 1) assessment, a midway (Level 2) assessment, and an end-of-unit (Level 3) assessment. Students not demonstrating mastery on Level 1 or Level 2 content could be retested within the content group cycle. Assessments could be administered immediately adjacent to instruction to help pinpoint areas where students needed additional support before proceeding to content on the next learning pathway Level. The PIE system is designed to allow this process to repeat throughout the school year with successive content groups.

### 2.2.1 Assessment and Instruction Cycles

The PIE system's instructionally embedded assessment module is designed to continuously evaluate student progress on the knowledge, skills, and understandings included in the learning pathways (LPs) aligned to fifth grade mathematics standards. Assessments are delivered directly after students receive instruction on LPs for relevant content standards. Teachers have flexibility to choose the composition of content standards and the timing of assessment delivery, while still adhering to local curriculum and pacing guidelines.

During the instructionally embedded assessment window, teachers customize discrete collections of LP-based content standards for instruction and assessment according to local curriculum and pacing guidelines. Teachers can prepare in advance by creating all of their content collections at the beginning of the academic year, or they can create these collections periodically as the year progresses. Either way, each unit of instruction and assessment is composed from one content collection at a time.

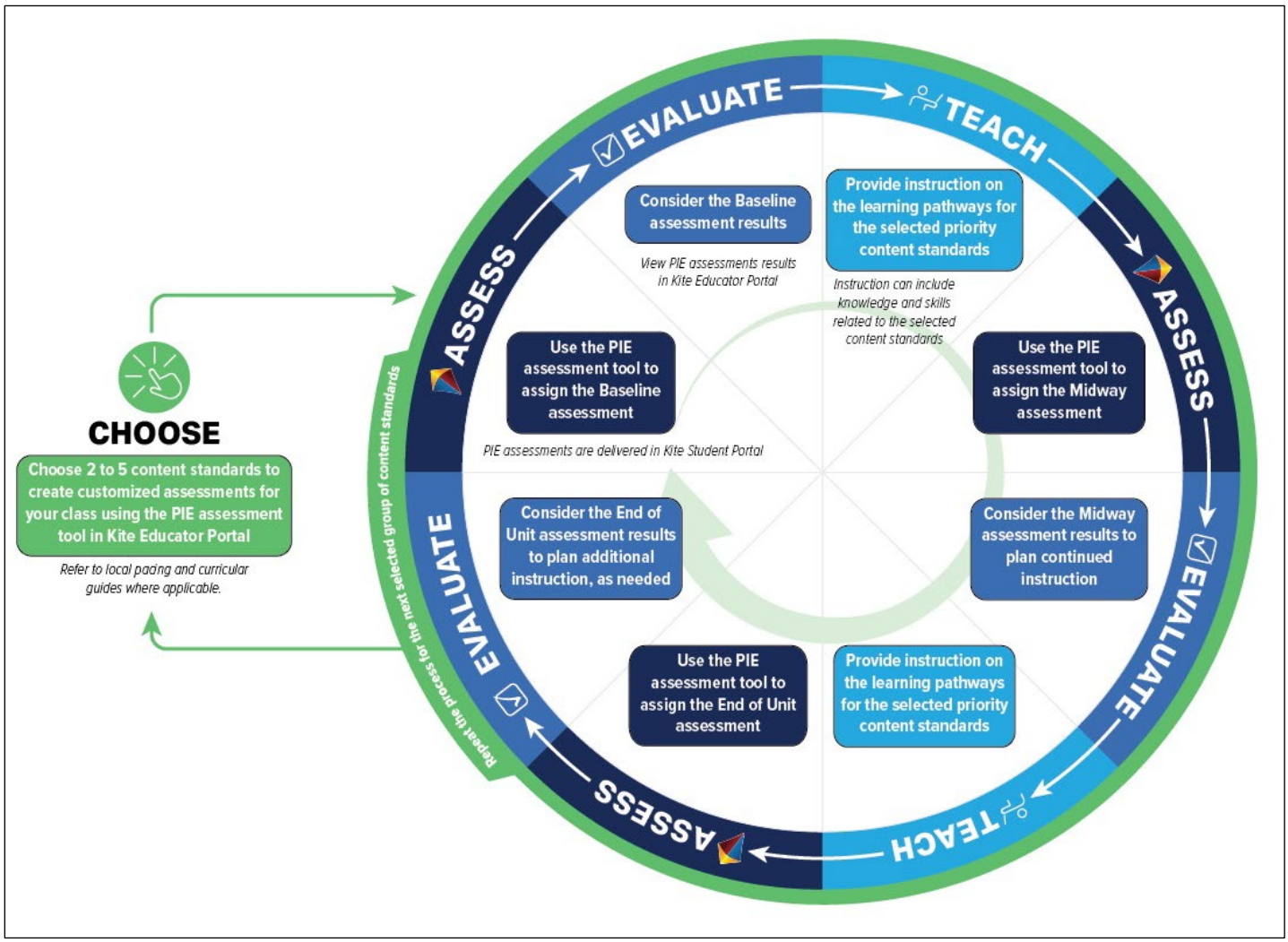
Teachers administer a short baseline assessment before they begin instruction on the first content collection they created. The baseline assessment comprises 6 to 15 items that measure Level 1 skills as defined by each LP for each of the selected standards in the content collection. In accordance with mastery learning approaches (Guskey, 2022), teachers use results from the baseline assessment to determine which students have not yet mastered any of the Level 1 skills needed to accomplish mastery of the grade-level skills. Teachers then begin whole-class instruction on the content collection with a focus on those Level 1 skills that are not yet mastered using supplemental small group and/or individual instruction as needed.

Teachers administer another short assessment halfway through the instructional unit that comprises another set of six to 15 items to measure Level 2 skills using the same collection of LPs. Three additional Level 1 items, one for each nonmastered Level 1 skill, are administered to students who did not master any of the Level 1 skills from the baseline assessment to evaluate progress toward mastery for these skills. Teachers use the progress results from the mid-unit assessment to inform instructional decisions and build in additional supports for the whole class, small group, or individual students who are still working toward mastery of prerequisite grade-level skills.

After the full instruction cycle has ended for a customized content collection, teachers deliver a short end-of-unit assessment similar in structure and length to the mid-unit assessment. This assessment measures target Level 3 skills and includes additional items to measure student progress on previously non-mastered Level 2 skills. Teachers then start a new instruction cycle with the next customized collection of LP-based content standards they created. The intended instruction and assessment cycle is illustrated in Figure 12.

Figure 12.

*The PIE Assessment and Instruction Cycle*



## 2.3 The Kite® Suite

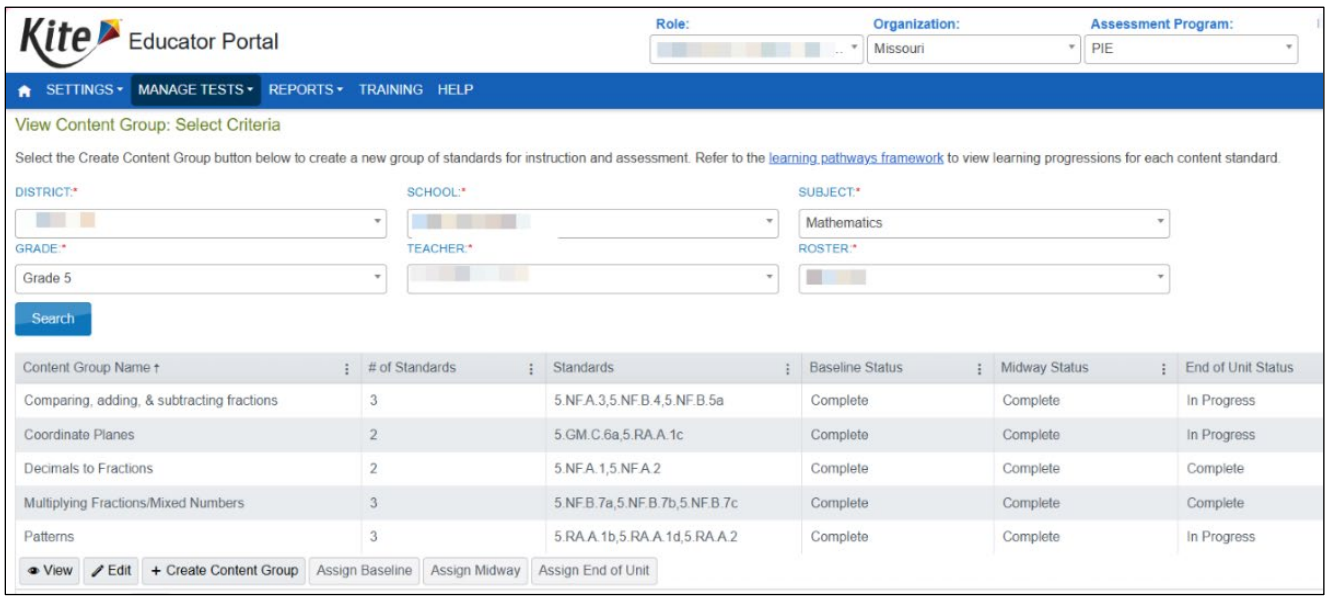
PIE assessments are managed and administered through the Kite® Suite digital platform. Teachers and test coordinators work within Kite Educator Portal to manage student data, create personalized assessments using teacher-selected content standards, administer assessments, and access assessment results. Students log in to Kite Student Portal to complete their assigned assessments.

In Kite Educator Portal, teachers created content groups to cover the full set of 25 available standards in the instructionally embedded window. Within the portal, teachers had the ability to create, edit, view, and delete content groups that had not yet been started. Each content group represents an assessment cycle, with three tests associated with it throughout the instructional process. Teachers could initiate these tests at their discretion by using the “Assign Baseline” (Level 1 items), “Assign Midway” (Level 2 items), and “Assign End of Unit” (Level 3 items) buttons. Each test was assigned a status, providing teachers with clear insights into whether the entire class had completed the assignment.

In December 2023, as part of pre-pilot focus group sessions, educators were shown a prototype version of the content group planning tool. PIE staff sought educator feedback on the ease of navigation of the tool, ease of creating content groupings, and any suggestions on training and resources to support the use of the planning tool. Educators offered some valuable suggestions, including a “Start Here” indicator and a linear path to follow, suggested wording for naming content groups and hints for assembling groups of standards, a way to generate a summary report for all content groups by date range, a way to view sample items by date range, and training checklists and/or videos. Feedback from these focus group sessions was incorporated into the final planner design used during the pilot study.

Figure 13.

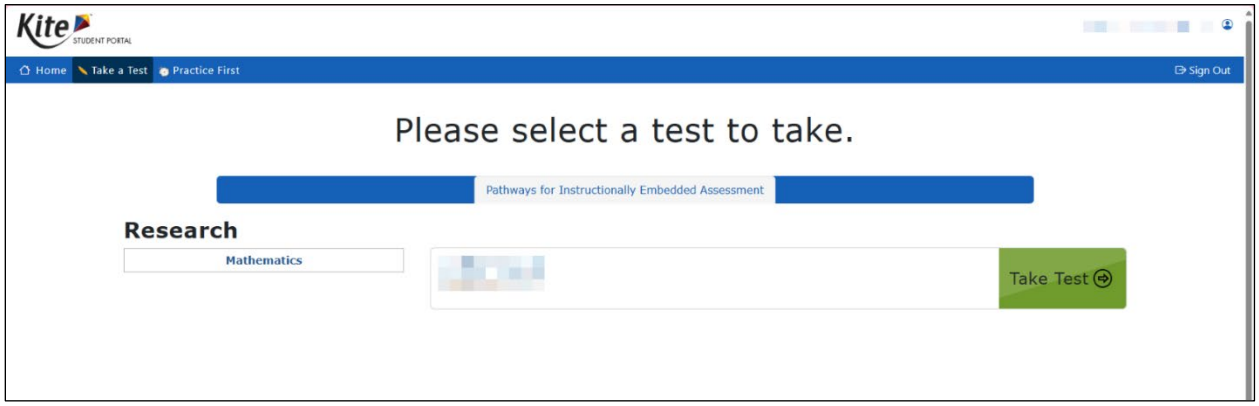
*PIE Content Group Planner Page in Kite Educator Portal*



Teachers were able to obtain the students' login credentials for Student Portal in Educator Portal Test Management. Teachers were asked to give the students their login credentials, allowing them to log in to Kite Student Portal to access the assessments. Clicking **Take Test** will launch the test.

Figure 14.

*Take Test Page in Kite Student Portal*



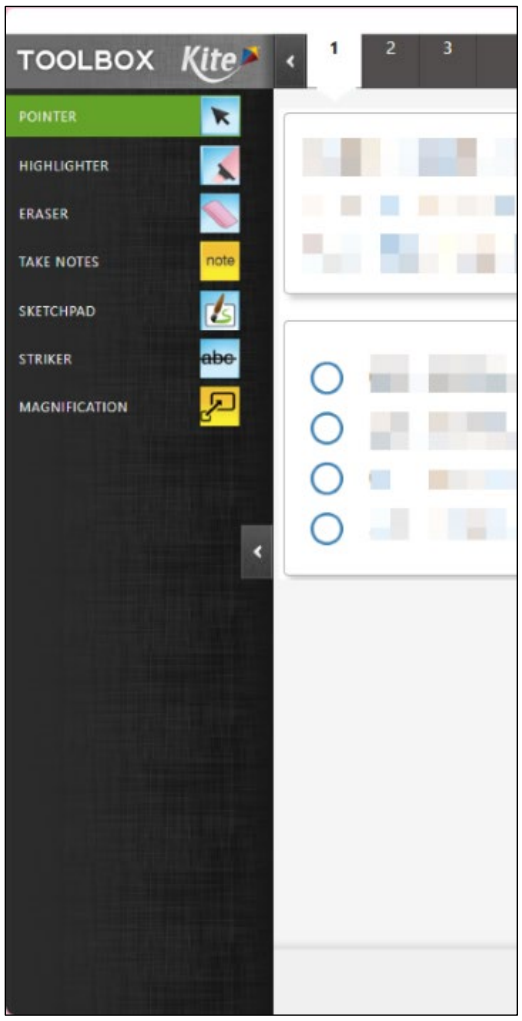

After each test begins, students are presented with a page that includes general directions and reminders. They can then click **BEGIN** to start the test. Students have access to a toolbox containing various tools, which are available on all tests and can assist them throughout the assessment. The tools provided for the PIE tests include: the **POINTER**, which allows students to switch back to the regular cursor; the **HIGHLIGHTER** and **ERASER**, which enable students to highlight important parts of an item or erase highlights; **MAGNIFICATION**, which enlarges the content on the screen; **NOTES**, which lets students take notes if needed; **SKETCHPAD**, which allows students to draw or sketch to help complete an item; and **STRIKER**, which lets students cross out answer options they believe are incorrect.

For additional support, the Kite Suite offers teachers the ability to enable Personal Needs and Preferences (PNP) options. These options, available for the PIE program, include: Color Contrast, Color Overlay, and Reverse Contrast, which adjust the page colors for easier readability; Masking, which allows students to cover parts of the screen; and Text-to-Speech (TTS), which reads the text on the screen aloud to students.

To navigate through the test, students can use a navigation bar at the top of the screen and navigation buttons at the bottom. Clicking on an item number at the top takes them directly to the selected item, while the **BACK** button allows them to return to the previous item. The **CLEAR** button removes any selected answer, and the **NEXT** button moves them to the next item. Finally, the **REVIEW/END** button takes students to the review-and-end screen.

Figure 15.

*PIE Toolbox Page in Kite Student Portal*




## 2.4 Final Assessment Delivery

After the close of the academic window, all participating students completed the final assessment, which was delivered in the Kite system as a single test form consisting of 50 items. The same fixed test form was administered to all students. The assessment was designed to measure Level 3 skills across 25 content standards, with two items administered for each standard. This ensured that all students were evaluated on the same set of items covering the full range of targeted content standards.

# Pilot Study Participation

Prior to the initiation of this pilot study, approval was obtained from the Institutional Review Board (IRB) at the University of Kansas (KU). This study was reviewed and approved under the protocol identified by IRB ID: STUDY001499737. The approved protocol outlines all procedures

related to participant recruitment, informed consent, data collection, data confidentiality, and risk mitigation measures.

## 3.1 Recruitment and Inclusion Criteria

MO-DESE agency staff were asked to recruit districts, schools, and fifth grade mathematics teachers to participate in the pilot study. Students receiving instruction on the Missouri fifth grade mathematics content standards were eligible to be included in the pilot study. Students with significant cognitive disabilities who take statewide alternate assessment were excluded from this study.

Expectations of each teacher participating in the PIE pilot study during the 2024–2025 school year included the following steps:

- Complete teacher training modules and professional learning activities.
- Create content groupings to be used as the basis for the assessment and instruction cycles during the PIE instructionally embedded assessment window.
- Help students log in and navigate to the Kite assessment software and monitor testing throughout the study.
- Administer the PIE technology practice test to students.
- Complete instruction and assessment cycles aligned to their content groupings during the instructionally embedded assessment window that was open from September 16, 2024, through February 21, 2025.
- Administer a final assessment during the final assessment window (February 24–March 21, 2025).
- Administer a survey to students regarding their experiences taking PIE assessments.
- Complete a survey about their experience using the PIE system.

PIE staff distributed recruitment messages to MO-DESE staff, who shared the recruitment message with districts in which schools met these criteria. Depending on MO-DESE's preferences, volunteer districts either notified the state or contacted PIE staff directly. After a district or school expressed interest, MO-DESE staff provided principals or other leadership with detailed information about participation and next steps. MO-DESE then sent a follow-up email to principals or relevant leaders to confirm their district's participation in the study. Afterward, districts or schools reached out to their teachers with a recruitment flyer that provided information on how teachers and students could participate.

Teachers indicated their interest in participating by responding to their principals. After teachers confirmed their willingness to participate, the school granted approval for their teachers and students to take part. After a district confirmed its participation, PIE staff sent an email to the school's test coordinators or main contacts that included a table outlining the key steps and timeline for participation (see Appendix A).

## 3.2 Participants

In total, 1,572 students participated and completed 19,117 test sessions during the instructionally embedded window. These test sessions were administered by 55 teachers across

28 districts, 32 schools, and 69 classroom rosters. Of the 28 school districts included, 19 were classified as rural, with the remaining evenly distributed among town, city, and suburban districts. On average, the participating schools had approximately 22 teachers on staff and 319 enrolled students. All students included in the analyses outlined in this technical report completed a full instruction and assessment cycle for at least one content standard. Of the 1,572 students who participated in the instructionally embedded assessment window, 1,187 students participated in the spring assessment.

Table 3 reports a breakdown of the demographic summary of the students that participated in the 2024–2025 PIE pilot study. The PIE student population was approximately balanced by gender. Approximately 77% of students were White, 15% were Black, and 6% were of two or more races. Students characterized as Asian, American Indian, Native Hawaiian or Pacific Islander, or Alaska Native comprised less than 2% of the combined student population. Approximately 93% of students were non-Hispanic, and approximately 95% of students were neither eligible for English-learning services or monitored.

Table 3.

*Demographic Summary of Students Participating in Pilot Study (N = 1,572)*

| Demographic group | *n* | % |
|---|---|---|
| Gender | | |
| Male | 790 | 50.3 |
| Female | 782 | 49.7 |
| Race | | |
| White | 1,218 | 77.5 |
| Black | 235 | 14.9 |
| Two or more races | 94 | 6.0 |
| Asian | 16 | 1.0 |
| American Indian | 5 | 0.3 |
| Native Hawaiian or Pacific Islander | 3 | 0.2 |
| Alaska Native | 1 | 0.1 |
| Ethnicity | | |
| Non-Hispanic | 1,472 | 93.6 |
| Hispanic | 100 | 6.4 |
| English-learning (EL) participation | | |
| Not EL eligible or monitored | 1,496 | 95.2 |
| EL eligible or monitored | 76 | 4.8 |

*Notes. N*= Number of students; % = Percentage of students

# Administration

The instructionally embedded window was September 13, 2024, through February 21, 2025, and the final assessment window was February 24–March 21, 2025. A teacher and student survey window was open from February 5–23, 2025.

## 4.1 Teacher Professional Learning

At the start of the school year, teachers participating in the pilot study attended virtual orientation meetings where they were provided with an overview of the pilot study activities, timeline, where to find test administration materials and an e-learning training course. After the orientation meetings, teachers completed a training course which comprised a series of six online modules that introduced the PIE learning pathways, demonstrated how to create content groupings, explained the PIE assessment design, provided a walk-through of the administration of PIE assessments, oriented teachers to PIE assessment reports, and provided an opportunity for teachers to interpret assessment data (for a mock class) and explore implications for classroom teaching via a model of data-based instructional decision-making. As a follow-up to the modules and to support learning transfer, teachers were invited to participate in office

hours hosted by state education leadership and were provided job aid resources that reinforced the module content.

The “double diamond” methodology for innovative design (Design Council, 2005) was employed to create a professional learning solution for teachers to develop proficiency in: use of learning pathways to create customized PIE assessments, administration of PIE assessments, and interpretation and use of PIE assessment data to inform classroom instruction. The methodology is characterized by a Discovery phase, in which the learning environment was analyzed; a Definition phase, in which a plan for a performance-based professional learning solution was proposed to stakeholders; a Development phase, in which end-user (teacher) input was collected via focus groups and incorporated in an iterative process; and the Delivery phase, in which teachers completed the self-paced learning and provided feedback on the experience (Nash et al., 2024).

The learning solution featured a series of six e-learning modules that introduced PIE system components, developed teacher assessment literacy, provided a walk-through of the administration of PIE, oriented teachers to PIE assessment reports, and provided an opportunity for teachers to interpret data (for a mock class) and explore implications for classroom teaching via a model of data-based instructional decision-making.

Within each module, a job aid was available which highlighted and summarized some of the key takeaways for teachers to have “at their fingertips.” Several one-page job aids were available within the learning solution:

- Example PIE Learning Pathway Map View
- PIE Assessment and Instruction Cycle (graphic organizer)
- How to Read a Learning Pathway Map
- How Test Length Relates to Your Content Groups
- How to Read PIE Assessment Reports
- Driving Your Instructional Decision-Making with Data

The professional learning solution was delivered through Moodle LMS (learning management system) using the Boost Union theme for accessibility and ease of navigation. Integrated through an LTI connection in the Educator Portal, the course allowed participants to complete the learning solution within the same platform used for administering the PIE assessment.

## 4.2 Resources and Manuals

Test administrators, school staff, and district staff were provided with multiple resources to support the assessment administration process. Resources were provided on the PIE website. To provide updates and reminders, participants were emailed a monthly newsletter addressing timely topics such as assessment deadlines, content group ideas from teachers and PIE team discussions, and tech tips.

This section provides an overview of resources and manuals available for test administrators and district-level staff on the PIE website. The PIE website provides specific resources and manuals designed for test administrators.

Table 4.

*PIE Resources for Test Administrators*

| Resource | Description |
|---|---|
| *PIE Learning Pathways* | Shows how students can move from the basic cognitive and early academic skills toward the knowledge, skills, and understandings they need to master the grade-level standards and help educators understand how those skills might logically follow one another |
| *PIE Pilot Test Administration Manual* | Describes test administrator responsibilities, allowable test administration practices, Kite Student Portal features and navigation, the technology practice test, scheduling and time limits, test security, test administration guidelines for teachers and students, and scripts for student directions for the instructionally embedded assessment and final assessment |
| *PIE Pilot Educator Portal User Guide* | Provides test administrators step-by-step procedures for using Kite Educator Portal for constructing content units by standards, assigning instructionally embedded assessments and accessing reports |
| *PIE Pilot Practice Test Directions* | Provides directions for students, parents, teachers, and others to navigate an assessment, learn about available tools and item types in Kite Student Portal prior to assessment |

The PIE website provides specific resources and manuals designed for three district-level supporting roles: assessment coordinator, data manager, and technology personnel.

Table 5.

*PIE Resources for District-Level Staff*

| Resource | Description |
|---|---|
| *PIE Pilot Technical Specifications Manual* | Provides technology personnel information and tools to prepare the network and devices for assessment administration |
| *PIE Pilot Test Coordinator Manual* | Provides test coordinators with information to support data management, data managers, technology personnel, and teachers |
| *PIE Pilot Educator Portal User Guide* | Provides data managers step-by-step procedures for using Kite Educator Portal for uploading user, enrollment, and roster data, as well as accessing reports |
| *PIE District Test Coordinator Training* | A video for test coordinators that provides an overview of the PIE assessment, responsibilities for different roles, test security, and important dates |
| *Roster Upload Template for Data Managers* | Provides the upload template for data managers to create rosters in Kite Educator Portal |
| *Enrollment Upload Template for Data Managers* | Provides the upload template for data managers to enroll students in Kite Educator Portal |

## 4.3 Security

Test security was promoted through the PIE required test administrator training and certification requirements for test administrators. Test administrators were expected to deliver PIE assessments with integrity and maintain the security of testlets. The training for assessment administration detailed test security measures. Test administrators completed their security agreement through Kite Educator Portal. Test administrators were not granted access to Kite Educator Portal if they had not completed the security agreement.

### Kite Test Security Agreement Text

The Kite Suite provides opportunities for flexible assessment administration; however, all assessments delivered during the school year are secure.

Test administrators and other educational staff who support implementation are responsible for following Kite test security standards:

1. Assessments and assessment items are not to be stored or saved on computers or personal storage devices; shared via email or other file sharing systems; or reproduced by any means.

2. Except where explicitly allowed as described in the *Test Administration Manual*, electronic materials used during assessment administration may not be printed.
3. Those who violate the Kite test security standards may be subject to their state's regulations or state education agency policy governing test security.
4. Users will not give out, loan or share their password with anyone. Allowing others access to an Educator Portal account may cause unauthorized access to private information. Access to educational records is governed by federal and state law.

Questions about security expectations should be directed to the local assessment coordinator.

## 4.4 Administration Evidence

One purpose of the pilot study was to evaluate the usability and feasibility of the PIE system in fifth grade mathematics, including administration of the assessment and instruction cycles, use of the progress reports and the training and resources designed to support implementation.

During the instructionally embedded window, teachers determined how many and which learning pathways aligned to the content standards to assess students on within an assessment and instruction cycle. The set of standards-aligned learning pathways a teacher selected for a cycle is referred to as a *content group*.

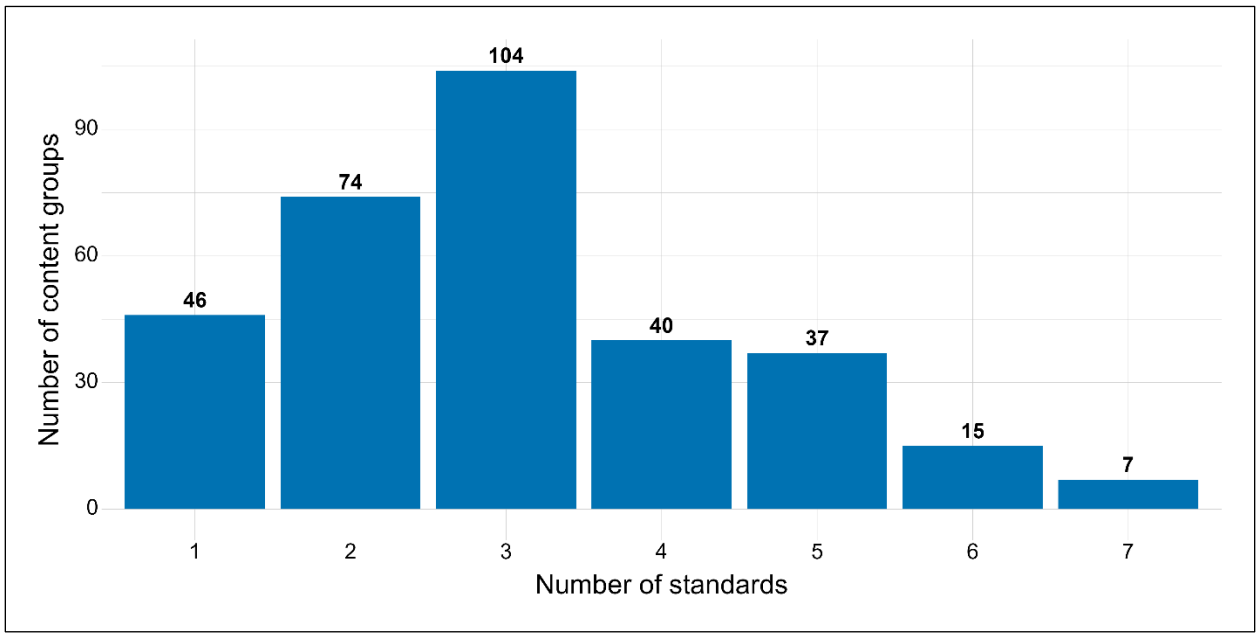
### 4.4.1 Number of Standards-Aligned Learning Pathways Selected for Content Groups

The assessment blueprint used for the 2024–2025 PIE instructionally embedded window covers 25 content standards (and their corresponding LPs) across four fifth-grade domains: number sense and operations in fractions, relationships and algebraic thinking, geometry and measurement, and data and statistics. It was recommended that content groups consist of between two and five standards; however, the Kite system allowed content groups to consist of between one and seven standards.

In total, 323 content groups were created by teachers in the Kite system during the instructionally embedded assessment window. Figure 16 provides a breakdown of the number of content standards selected for the content groups created by teachers. Two hundred fifty-five (78.9%) content groups were created to assess students in an adequate number of standards, where an adequate number of standards was determined as between two and five standards. In contrast, 68 (21.1%) content groups were created to assess students in an inadequate number of standards, defined in this context as a content group containing one standard or six or more standards.

Figure 16.

*Number of Standards Included in a Content Group*




## 4.4.2 Patterns in Standards-Aligned LPs Selected for Content Groups

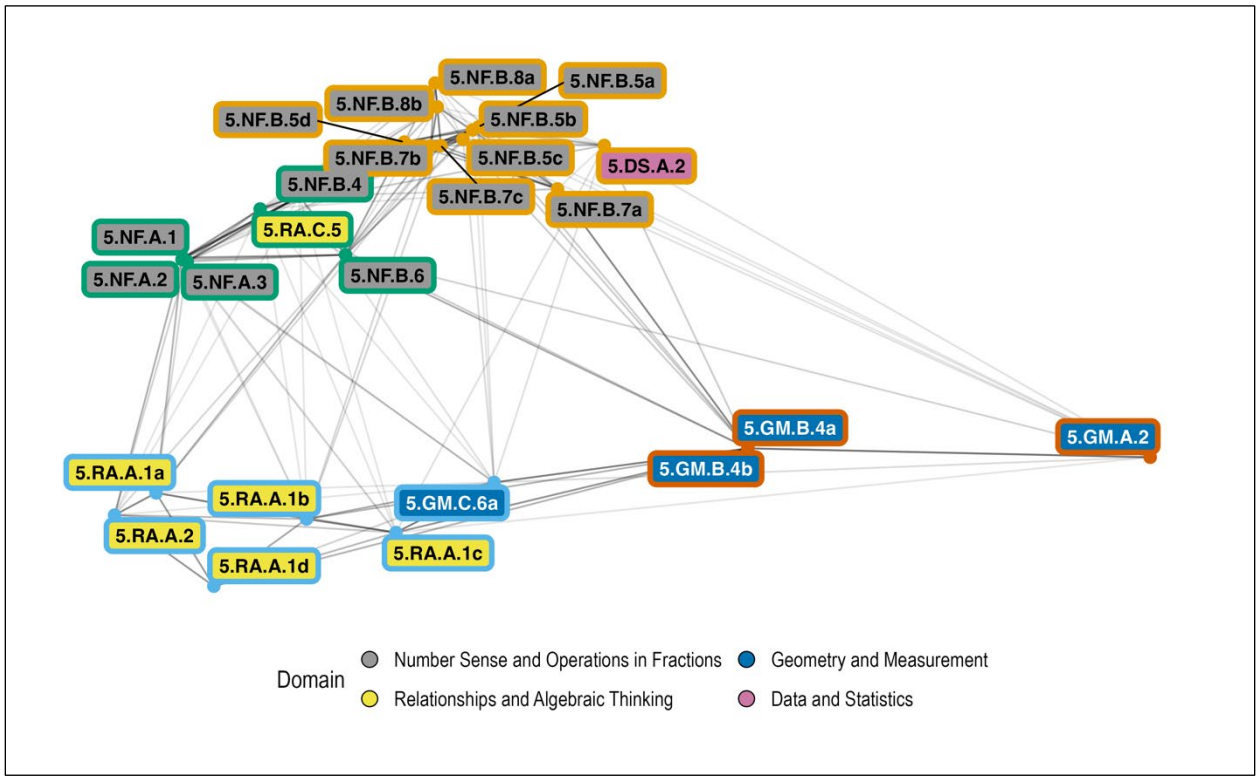
Although teachers were provided guidelines on the number of standards-aligned LPs to select for each content group, teachers could determine which standards would be grouped together for instruction. To explore the ways in which teachers selected standards-aligned LPs for their content groups, we conducted a network analysis to identify clusters of standards that were often selected together for instruction, and each cluster of standards is referred to as a *collection*.

Figure 17 depicts a network graph of standards, along with collections that were identified within the network. The collections are depicted as the outline color surrounding the label for each standard. Preliminary results suggest that the standards form four main collections, and the standards map to each of the four collections largely according to the content domain. The first collection at the top of Figure 17 (outlined in orange) consists mainly of standards from strand B of the number sense and operations in fractions domain. The one standard from the data and statistics domain is also included in this collection. The second collection on the left-hand side of Figure 17 (outlined in green) consists of the standards from strand A of the number sense and operations in fractions domain, along with the one standard from strand C of the relationships and algebraic thinking domain and the two remaining standards from strand B of the number sense and operations in fractions domain. The third collection at the bottom of Figure 17 (outlined in light blue) consists of standards from strand A of the relationships and algebraic thinking domain along with the one standard from strand C of the geometry and measurement domain. Finally, the fourth collection on the right-hand side of Figure 17 (outlined in red) consists of the remaining standards from the geometry and measurement domain.

These collections are largely consistent with the intended uses of the Instructionally Embedded model. Evidence that collections consist of standards from the same domain indicates that teachers selected related content to be assessed together. However, evidence that collections do not perfectly align to content domains also indicates that teachers made content judgments about which standards should be taught and assessed together, beyond domain and strand information, likely informed by local curriculum as well as the learning pathways underlying each content standard.

Figure 17.

*Network Graph Depicting Common Groupings of Standards*




## 4.4.3 Sequence of Content Standard Selection

While teachers were provided guidance on the number of content standards to select for content groups, teachers were given flexibility to determine the sequence in which content standards were selected for instruction.

Table 6 describes patterns in the sequence in which standards were selected by teachers for the content groups in Kite Educator Portal. For each standard, the mean and median instructional cycle in which a standard was administered by teachers is reported (e.g., first group, second group, etc.). Further, we report variability in administration across instructional cycles via a standard deviation. The findings indicates that teachers tended to administer standards belonging to Collection 2 during the first two instructional cycles of the academic year. Standards belonging to Collections 3 and 4 were typically assessed during the second and third

instructional cycles. Last, standards identified in Collection 1 were typically assessed in the third or fourth instructional cycle.

Table 6.

*Summary of Content Standard Selection by Sequence of Instruction and Assessment Cycles*

| Content standard | No. of cycles | | | | |
|---|---|---|---|---|---|
| | *Mean cycle #* | *Median cycle #* | *Standard Deviation* | *Minimum cycle #* | *Maximum cycle #* |
| Collection 1 | | | | | |
| 5.NF.B.7b | 3.8 | 3.0 | 1.30 | 2 | 9 |
| 5.NF.B.7c | 3.6 | 3.0 | 1.37 | 2 | 9 |
| 5.NF.B.8b | 4.2 | 3.0 | 1.87 | 2 | 10 |
| 5.NF.B.5b | 3.8 | 3.5 | 1.44 | 2 | 8 |
| 5.DS.A.2 | 4.8 | 4.0 | 2.82 | 2 | 9 |
| 5.NF.B.5a | 4.0 | 4.0 | 1.31 | 2 | 8 |
| 5.NF.B.5c | 3.8 | 4.0 | 1.11 | 2 | 8 |
| 5.NF.B.5d | 3.4 | 4.0 | 1.53 | 1 | 8 |
| 5.NF.B.7a | 4.6 | 4.0 | 2.22 | 1 | 9 |
| 5.NF.B.8a | 4.4 | 4.0 | 1.95 | 2 | 10 |
| Collection 2 | | | | | |
| 5.NF.A.1 | 1.7 | 1.0 | 1.26 | 1 | 6 |
| 5.NF.A.2 | 1.9 | 1.0 | 1.40 | 1 | 6 |
| 5.NF.A.3 | 1.9 | 1.0 | 1.40 | 1 | 6 |
| 5.NF.B.4 | 2.3 | 2.0 | 1.27 | 1 | 6 |
| 5.NF.B.6 | 2.4 | 2.0 | 1.15 | 1 | 5 |
| 5.RA.C.5 | 4.0 | 2.0 | 3.35 | 1 | 9 |
| Collection 3 | | | | | |
| 5.RA.A.1c | 3.0 | 2.0 | 1.75 | 1 | 7 |
| 5.GM.C.6a | 3.1 | 3.0 | 1.66 | 1 | 7 |
| 5.RA.A.1a | 3.5 | 3.0 | 1.46 | 1 | 6 |
| 5.RA.A.1b | 3.2 | 3.0 | 1.78 | 1 | 6 |
| 5.RA.A.1d | 3.8 | 5.0 | 1.73 | 1 | 6 |
| 5.RA.A.2 | 3.9 | 5.0 | 1.47 | 1 | 6 |
| Collection 4 | | | | | |
| 5.GM.B.4a | 3.3 | 2.0 | 2.43 | 1 | 8 |
| 5.GM.B.4b | 3.3 | 2.0 | 2.43 | 1 | 8 |
| 5.GM.A.2 | 3.6 | 3.0 | 2.16 | 1 | 8 |

### 4.4.4 Pacing of Assessment and Instruction Cycles

To understand how teachers paced assessment administration across the instructionally embedded window, we analyzed timestamps indicating when classroom rosters began administering each content standard. The administration data were analyzed at the individual

content standard level (see Figure 18) and aggregated across standards (see Figure 19). In general, we would expect to see relatively consistent administration across the window, suggesting that an appropriate amount of instructional time is allocated according to the size of each content group. However, variations are expected due to several factors, including local pacing guides, the alignment of Missouri content standards with local curricula, and time spent at the beginning of the year on review.

Figure 18 displays the number of classroom rosters initiating instruction on individual content standards, by month. Patterns vary widely across standards, but several standards show high concentrations of administration in September or January, with declines in activity from October through December. This suggests substantial variation in instructional sequencing choices across classrooms, likely influenced by local curriculum pacing, perceived complexity of content standards, or strategic planning around breaks in instruction (i.e., winter break).

Figure 19 aggregates this information across standards and provides a high-level summary of pacing trends. Administration volume peaked in September and gradually declined through the fall months, reaching a low in December. A notable increase in administration occurred in January, more than twice that of any earlier month, before declining in February. These findings suggest a common administration pattern in which some standards were emphasized early in the instructionally embedded window, which was followed by a decline in administration activity heading into the winter break and, to meet coverage requirements, a noticeable acceleration in instruction after the winter break.

Notably, all four clusters of content standards (color coded in Figure 18) followed this general pattern, though their relative volume and timing differed. Cluster 2 (green), for example, contributed heavily to September's activity, while Cluster 3 (blue) and Cluster 1 (golden yellow) were the primary contributors to the increase in January. These trends point to differences not only in when standards were instructed, but also in how content was grouped and prioritized within instructional planning.

Figure 18.

*Pacing of Learning Pathway Instruction and Assessment, by Standard*

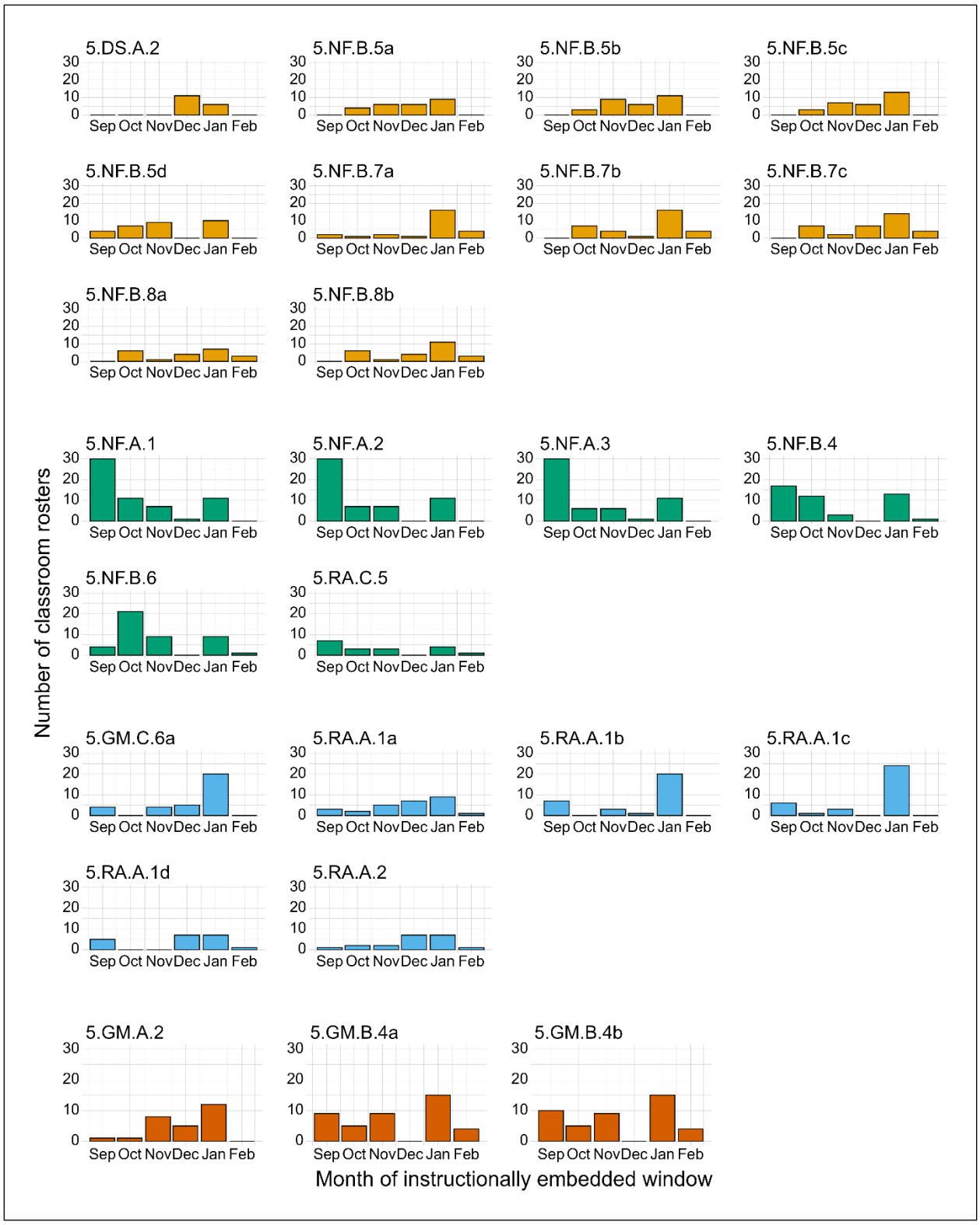
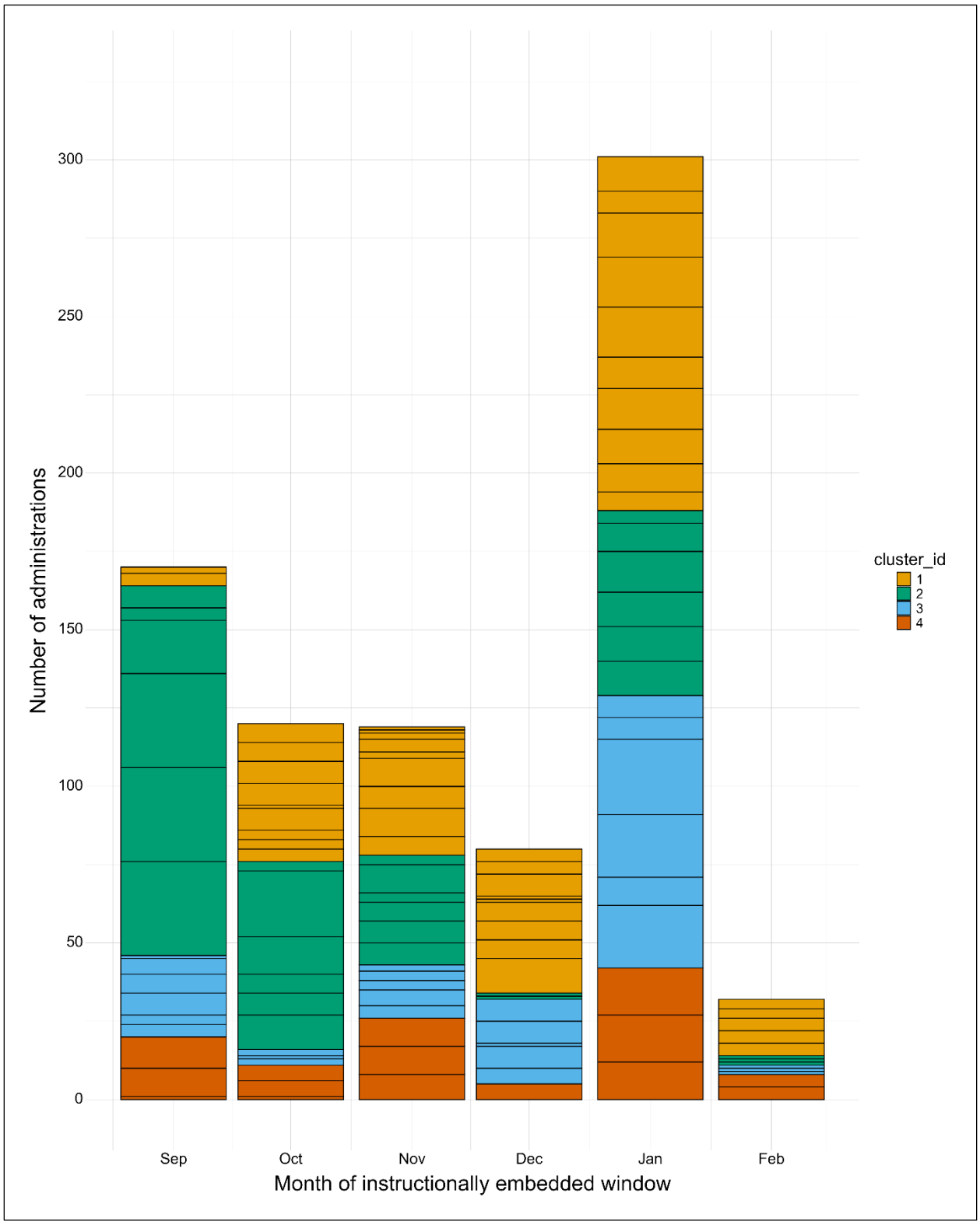


Figure 19.

*Pacing of Assessment Aggregated Across Learning Pathways*



Patterns of assessment pacing can be further studied at the student level by examining test session completions throughout the instructionally embedded window. Figure 20 shows the number of test sessions students completed, by week of the administration window. The student test forms demonstrated a consistent pattern of completion from the beginning of the window in September through the winter break in December. However, there was a noticeable spike in the number of test forms completed in January and February.

Figure 20.

*Distribution of Students' Completed Test Forms, by Week of Administration Window*

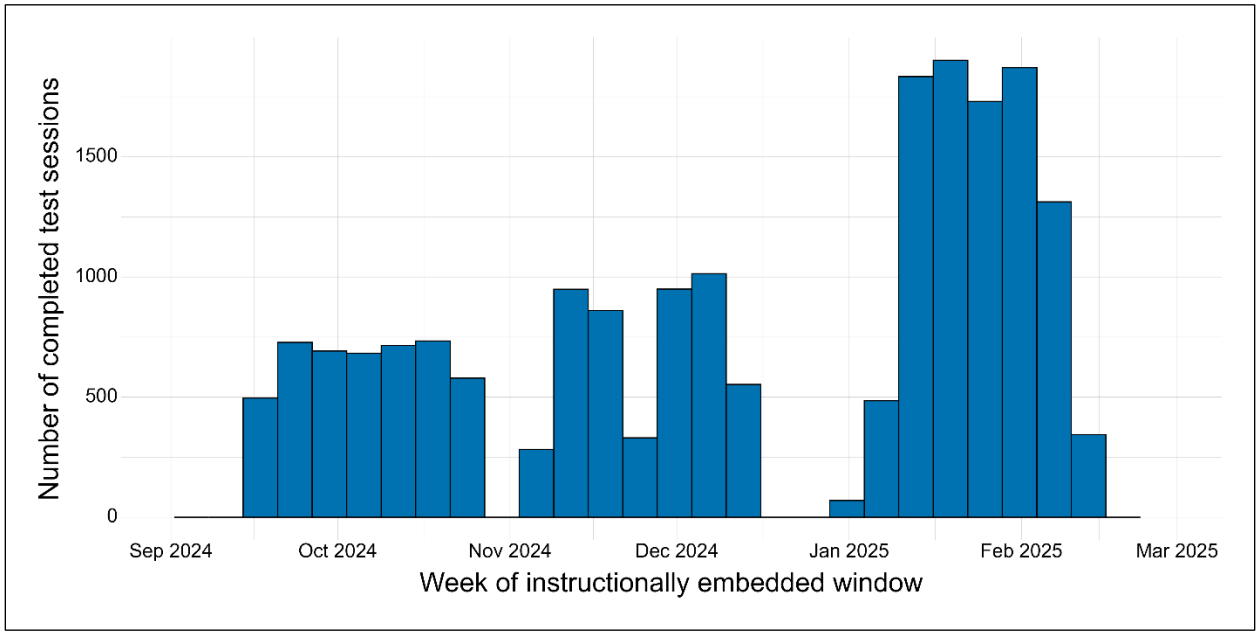
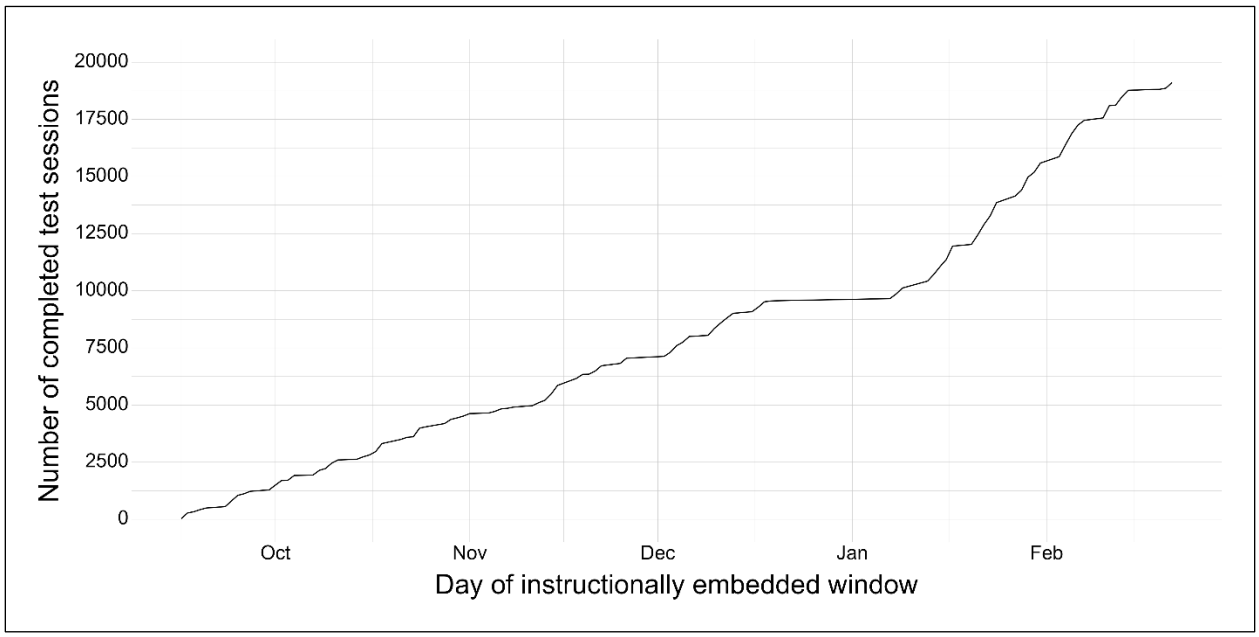



Figure 21 illustrates the timeline of total test form completion across all students. The figure shows the total test forms completed by each week of the instructionally embedded window. The cumulative pattern shows a consistent trend in test form completions through the winter break in December, followed by a surge in test form completion in January and February. In fact, about half of the test forms were completed after the winter break.

Figure 21.

*Cumulative Distribution of Completed Test Sessions Across All Students*




## 4.4.5 Assessment Completion Patterns

The PIE instructionally embedded model was intended to give teachers flexibility in the selection and pacing of assessments for content standards outlined in the assessment blueprint. Instructionally embedded assessments were available for 25 content standards-aligned LPs outlined in the assessment blueprint.

Figure 22 reports a breakdown of assessment blueprint coverage at the classroom roster level. Across the window, 11 (15.9%) classrooms initiated between one and five standards, 13 (18.8%) classrooms initiated between six and 10 standards, 22 (31.9%) classrooms initiated between 11 and 15 standards, 15 (21.7%) classrooms initiated between 16 and 20 standards, and 8 (11.6%) classrooms initiated between 21 and 25 standards. In total, 14 (20.3%) classrooms completed between one and five standards, 20 (29%) classrooms completed between six and 10 standards, 15 (20.7%) classrooms completed between 11 and 15 standards, 12 (17.4%) classrooms completed between 16 and 20 standards, and eight (11.6%) classrooms completed between 21 and 25 standards. However, only one (1.4%) classroom completed all 25 standards outlined in the assessment blueprint.

Figure 22.

*Assessment Blueprint Coverage Across Classroom Rosters*

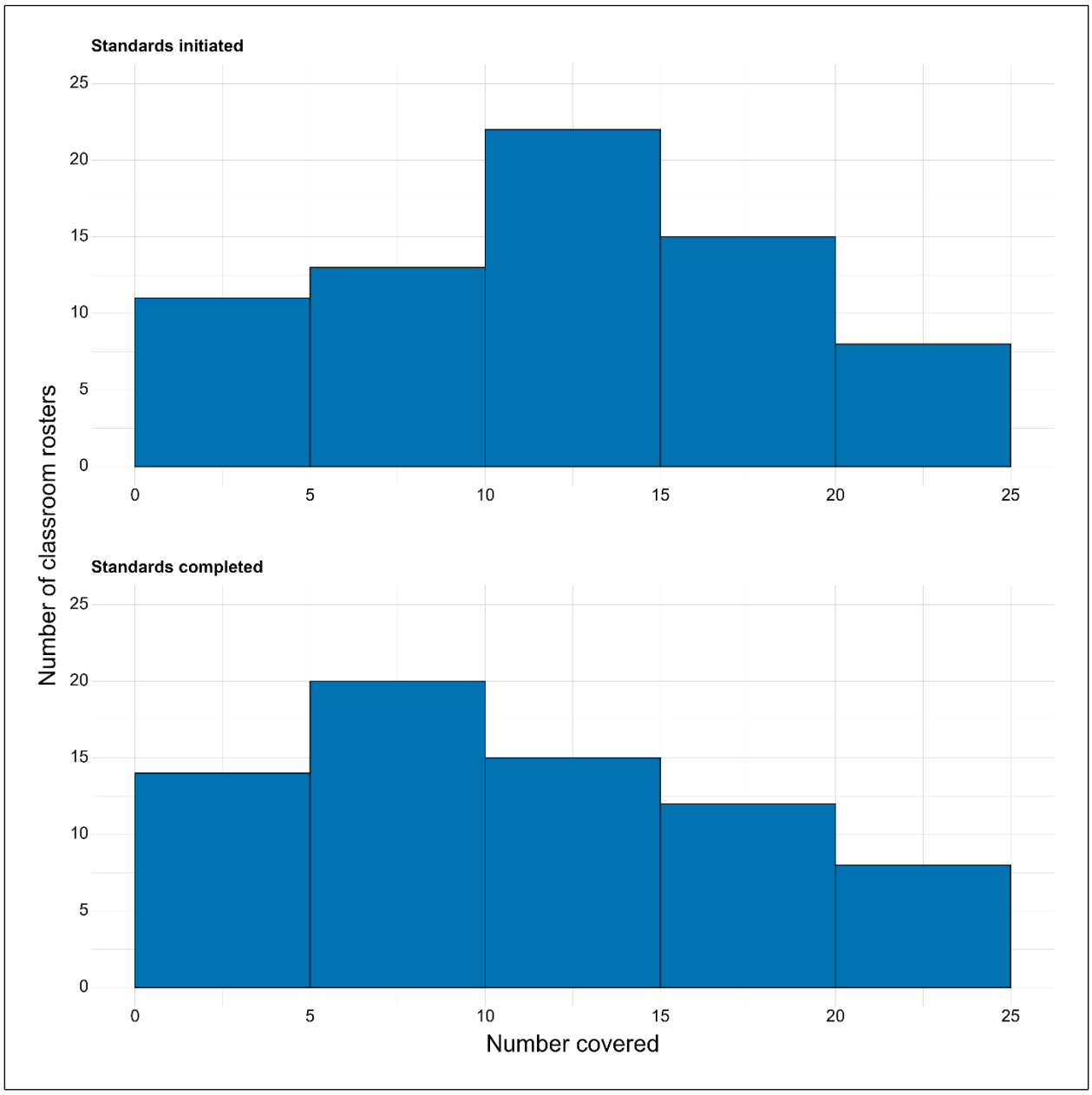
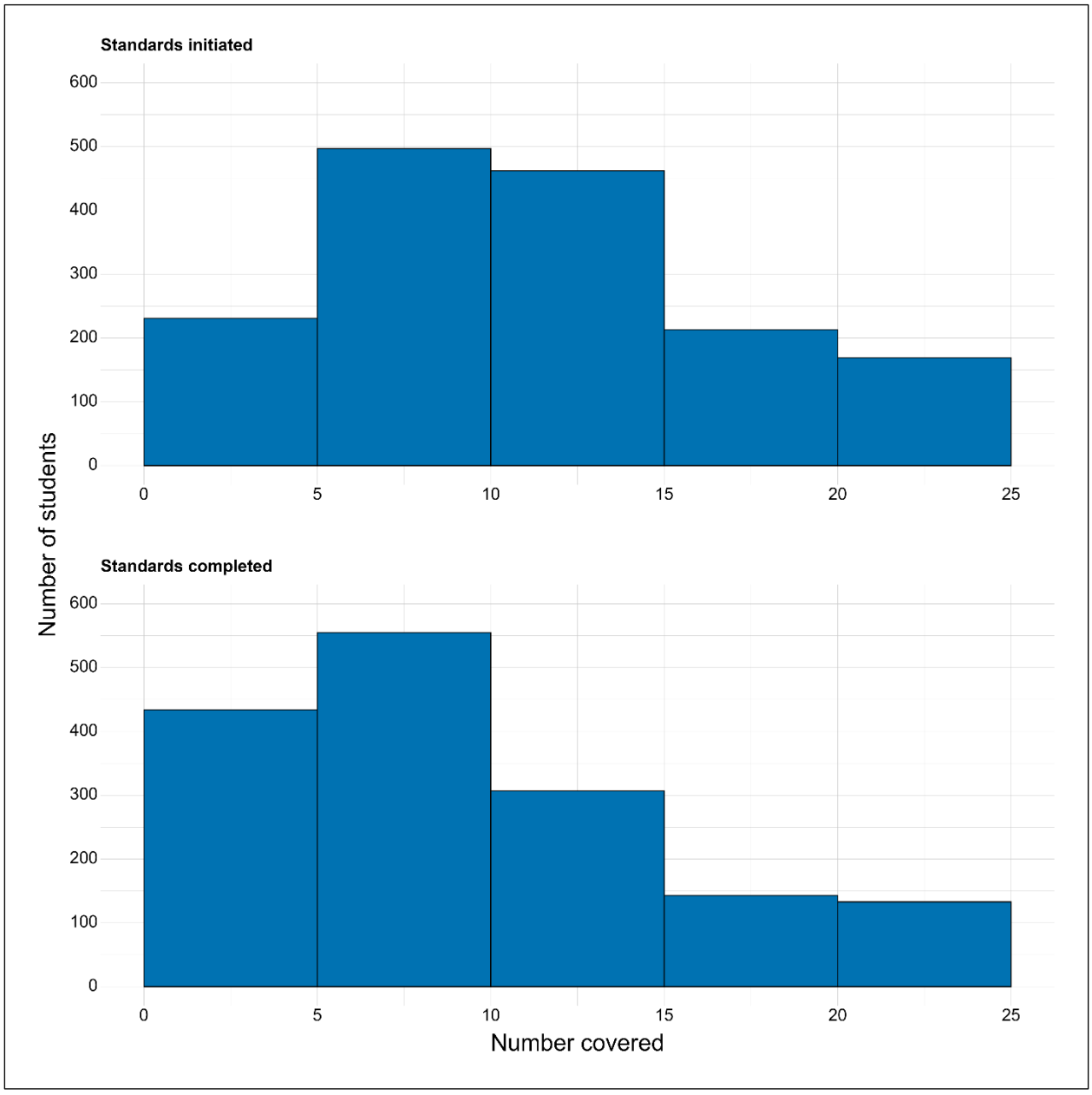



Figure 23 reports a breakdown of assessment blueprint coverage at the student level. Across the window, 231 (14.7%) students initiated between one and five standards, 497 (31.6%) students initiated between six and 10 standards, 462 (29.4%) students initiated between 11 and 15 standards, 213 (13.5%) students initiated between 16 and 20 standards, and 169 (10.8%) students initiated between 21 and 25 standards. In total, 484 (27.6%) students completed between one and five standards, 555 (35.3%) students completed between six and 10 standards, 307 (19.5%) students completed between 11 and 15 standards, 143 (9.1%) students completed between 16 and 20 standards, and 133 (8.5%) students completed initiated between

21 and 25 standards. However, only 26 (1.7%) students completed all 25 standards outlined in the assessment blueprint.

Figure 23.

*Assessment Blueprint Coverage Across Students*




## 4.4.6 Administration Timing

Figure 24 shows the distribution of time elapsed between when teachers began administering assessments within instruction and assessment cycles during the instructionally embedded window. We report the number of days between assessments as a proxy for the amount of time teachers may have used to provide instruction. Further, we provide a breakdown of instruction

time between individual assessments to shed light on the amount of instruction provided for: Level 2 concepts assessed at the midway assessment (and Level 1 concepts that were not mastered at the baseline assessment); and Level 3 concepts assessed at the end-of-unit assessment (and Level 2 concepts that were not mastered at the midway assessment). The findings show that teachers were consistent in their timing of assessment administrations.

The time allowed for instruction between assessments was approximately balanced between baseline and midway assessments and between midway and end-of-unit assessments. Teachers spent 1 week (1–5 days) instructing on Level 2 concepts (and Level 1 skills that required retesting) prior to the midway assessment in 54.1% of the cycles and 1 week (1–5 days) instructing on Level 3 concepts (and Level 2 skills that required retesting) prior to the end-of-unit assessment in 55.6% of the cycles. Teachers spent 2 weeks (6–10 days) instructing on Level 2 concepts (and Level 1 skills that required retesting) prior to the midway assessment in 25.6% of the cycles and 2 weeks (6–10 days) instructing on Level 3 concepts (and Level 2 skills that required retesting) prior to the end-of-unit assessment in 24.8% of the cycles. Last, teachers spent three or more weeks (11 or more days) instructing on Level 2 concepts (and Level 1 skills that required retesting) prior to the midway assessment in 20.3% of the cycles and three or more weeks (11 or more days) instructing on Level 3 concepts (and Level 2 skills that required retesting) prior to the end-of-unit assessment in 19.6% of the cycles.

Figure 24.

*Distribution of Time Elapsed Between Teachers' Assessment Administrations*

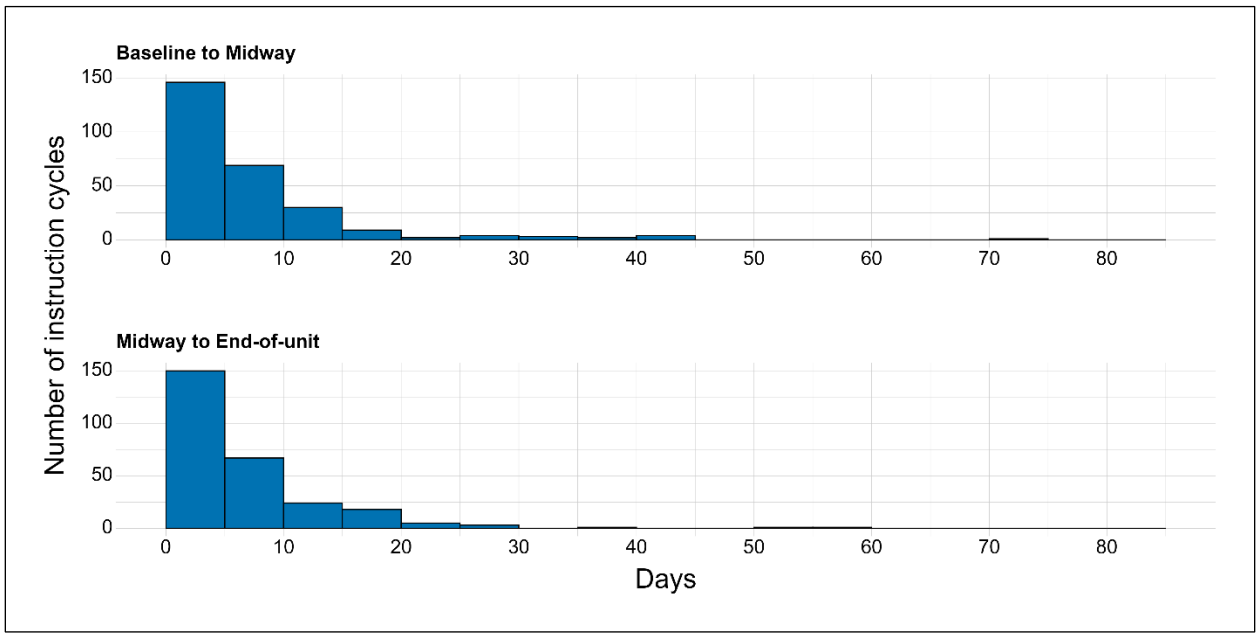



Table 7 further reports a summary of the time elapsed between teachers' assessment administrations, by number of standards assessed in each cycle. In general, we would expect more instruction time to be embedded into instruction and assessment cycles when addressing a higher volume of content standards. We observed little variability in the amount of time allocated for instruction for cycles addressing between one and four content standards, with the

median number of days allocated for instruction ranging from 5 to 6 days for Level 2 skills assessed at the midway assessment (and Level 1 skills not mastered at the baseline assessment) and ranging from 5 to 6 days for Level 3 skills assessed at the end-of-unit assessment (and Level 2 skills not mastered at the midway assessment). For cycles containing between five and seven content standards, the median number of days allocated for instruction ranged from 10 to 31 days for Level 2 skills assessed at the midway assessment (and Level 1 skills not mastered at the baseline assessment) and ranging from 10 to 18 days for Level 3 skills assessed at the end-of-unit assessment (and Level 2 skills not mastered at the midway assessment). However, the findings suggest a sharp increase in time elapsed between assessments for cycles containing six standards before sharply decreasing for cycles containing seven standards. This finding suggests that teachers may not have adequately allocated enough instruction time during cycles containing a volume of standards that exceeds the recommendations outlined in the training resources (e.g., two to five standards) provided to teachers prior to beginning the instructionally embedded window.

Table 7.

*Summary of Days Elapsed Between Teachers' Assessment Administrations, by Number of Standards Assessed*

| No. of standards in content group | *n* | *Min* | *Mdn* | *M* | *Max* | 25Q | 75Q | IQR |
|---|---|---|---|---|---|---|---|---|
| Baseline to midway | | | | | | | | |
| 1 | 41 | 2 | 5.0 | 6.0 | 18 | 3 | 8.00 | 5.00 |
| 2 | 54 | 2 | 5.5 | 6.7 | 41 | 4 | 8.00 | 4.00 |
| 3 | 96 | 2 | 5.0 | 8.1 | 74 | 4 | 9.25 | 5.25 |
| 4 | 32 | 2 | 6.0 | 7.8 | 20 | 5 | 9.25 | 4.25 |
| 5 | 29 | 2 | 10.0 | 11.3 | 42 | 7 | 14.00 | 7.00 |
| 6 | 14 | 5 | 31.0 | 26.5 | 42 | 16 | 41.00 | 25.00 |
| 7 | 4 | 11 | 11.0 | 11.2 | 12 | 11 | 11.25 | 0.25 |
| Midway to end of unit | | | | | | | | |
| 1 | 41 | 2 | 5 | 5.8 | 15 | 3 | 7.00 | 4.00 |
| 2 | 54 | 2 | 5 | 7.0 | 25 | 3 | 10.00 | 7.00 |
| 3 | 96 | 2 | 6 | 8.1 | 61 | 3 | 8.25 | 5.25 |
| 4 | 32 | 3 | 6 | 7.8 | 38 | 5 | 7.00 | 2.00 |
| 5 | 29 | 2 | 8 | 9.9 | 20 | 7 | 15.00 | 8.00 |
| 6 | 14 | 3 | 18 | 18.1 | 28 | 16 | 26.00 | 10.00 |
| 7 | 4 | 9 | 9 | 9.0 | 9 | 9 | 9.00 | 0.00 |

*Note.* 25Q = lower quartile; 75Q = upper quartile; IQR = interquartile range.

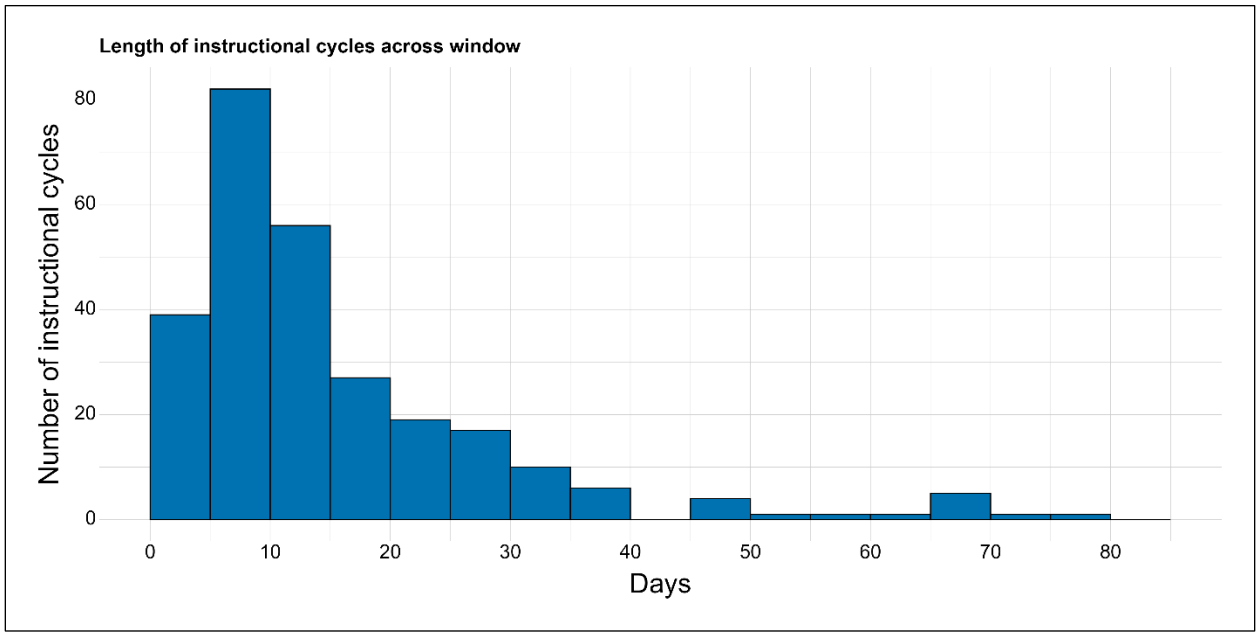
### 4.4.7 Number of Days Used to Complete Instructional Cycles

During the 2024–2025 academic year, teachers were expected to create instructional plans, instruct on these concepts, and assess students at a consistent pace. Informed by the training and resources provided in advance, teachers were expected to complete instructional cycles within a recommended 2-to-4-week time frame.

Figure 25 shows the distribution of instructional cycle lengths across the window. Thirty-eight (14.1%) cycles were completed within 5 school days (i.e., 1 week). Eighty-three (30.7%) cycles were completed between 6 and 10 school days (i.e., 2 weeks). Fifty-six (20.7%) cycles were completed between 11 and 15 school days (i.e., 3 weeks). Twenty-seven (10.0%) cycles were completed between 16 and 20 school days (i.e., 4 weeks). Last, sixty-six (24.4%) assessment and instructional cycles were completed in more than 20 school days (i.e., 5 or more weeks).

Figure 25.

*Distribution of Number of Days Used to Complete Instructional Cycles*




Throughout the window, certain cycles required more time to complete than recommended. In 30 (11.1%) instructional cycles, classrooms used more than 30 days (6 weeks) to complete the cycle. In 14 (5.2%) instructional cycles, classrooms used more than 40 days (8 weeks) to complete the cycle. Therefore, it may be helpful to examine conditional summaries of the number of days used to complete cycles, by classroom blueprint coverage and by content group size. Further, in the presence of outliers, the median number of days used to complete a cycle is a more robust representation compared to the mean number of days used to complete a cycle.

Table 8 shows the distribution of days used to complete instructional cycles based on the number of standards selected for a content group. The median number of days used to complete a cycle was less than 20 days (4 weeks) for all content group sizes, except for groupings that included six standards. In general, a vast majority of classroom rosters that selected the recommended number of standards per content group (i.e., two to five standards)

completed the cycle within 2 to 4 weeks. However, classrooms that exceeded five standards demonstrated substantially longer timeframes for completion of a cycle. In particular, the data shows that over 75% of the classrooms that administered content groups containing six standards used seven or more weeks to complete the cycle. These findings suggest that teachers who exceeded the recommended number of standards when selecting content groups for an instructional cycle were less efficient in completing an instructional cycle.

Table 8.

*Distribution of Days Used to Complete Instruction and Assessment Cycles*

| No. of standards in instructional cycle | *n* | *Min* | *Mdn* | *M* | *Max* | 25Q | 75Q | IQR |
|---|---|---|---|---|---|---|---|---|
| 1 | 41 | 3 | 10 | 10.8 | 27 | 5 | 12.00 | 7.00 |
| 2 | 54 | 3 | 10 | 12.7 | 49 | 7 | 17.00 | 10.00 |
| 3 | 96 | 3 | 10 | 15.3 | 80 | 7 | 20.00 | 13.00 |
| 4 | 32 | 4 | 11 | 14.7 | 57 | 9 | 15.00 | 6.00 |
| 5 | 29 | 5 | 17 | 20.2 | 61 | 13 | 24.00 | 11.00 |
| 6 | 14 | 7 | 48 | 43.6 | 69 | 31 | 67.00 | 36.00 |
| 7 | 4 | 19 | 19 | 19.2 | 20 | 19 | 19.25 | 0.25 |

*Note. n* = Number of cycles; Min = Minimum; Max = Maximum; 25Q = lower quartile; 75Q = upper quartile; IQR = interquartile range.

### 4.4.8 Multiple Opportunities to Demonstrate Mastery of Pathway Levels

During the instructionally embedded window, teachers are expected to reassess students on skills aligned to pathway levels that were not mastered during their initial assessment. For instance, the midway assessment provides an opportunity for teachers to reassess mastery of a Level 1 skill in cases where a student did not master the Level 1 skill at the baseline assessment. Further, the end-of-unit assessment provides an opportunity for teachers to reassess mastery of a Level 2 skill in cases where a student did not master the Level 2 skill at the midway assessment.

Table 9 reports the total number of students that tested on each standard and the number (and percentage) of students that retested on Level 1 and Level 2 skills within each standard. For Level 1 skills, the percentage of students who retested ranged from approximately 5.0% to 78.7%. For Level 2 skills, the percentage of students who retested ranged from approximately 9.7% to 83.0%.

Table 9.

*Patterns of Retesting in Math Content Standards*

| Content standard | *n* | Level 1 | Level 2 |
|---|---|---|---|
| 5.NF.B.7b | 420 | 29 (6.9%) | 151 (36.0%) |
| 5.NF.B.7c | 457 | 26 (5.7%) | 87 (19.0%) |
| 5.NF.B.8b | 367 | 96 (26.2%) | 212 (57.8%) |
| 5.NF.B.5b | 464 | 183 (39.4%) | 191 (41.2%) |
| 5.DS.A2 | 373 | 144 (38.6%) | 258 (69.2%) |
| 5.NF.B.5a | 372 | 29 (7.8%) | 118 (31.7%) |
| 5.NF.B.5c | 439 | 232 (52.8%) | 287 (65.4%) |
| 5.NF.B.5d | 583 | 459 (78.7%) | 323 (55.4%) |
| 5.NF.B.7a | 393 | 21 (5.3%) | 160 (40.7%) |
| 5.NF.B.8a | 318 | 16 (5.0%) | 205 (64.5%) |
| 5.NF.A.1 | 1,139 | 189 (16.6%) | 256 (22.5%) |
| 5.NF.A.2 | 948 | 545 (57.5%) | 574 (60.5%) |
| 5.NF.A.3 | 878 | 442 (50.3%) | 473 (53.9%) |
| 5.NF.B.4 | 722 | 333 (46.1%) | 348 (48.2%) |
| 5.NF.B.6 | 837 | 163 (19.5%) | 468 (55.9%) |
| 5.RA.C.5 | 296 | 130 (43.9%) | 154 (52.0%) |
| 5.RA.A.1c | 729 | 227 (31.1%) | 307 (42.1%) |
| 5.GM.C.6a | 648 | 209 (32.3%) | 344 (53.1%) |
| 5.RA.A.1a | 503 | 220 (43.7%) | 242 (48.1%) |
| 5.RA.A.1b | 659 | 98 (14.9%) | 64 (9.7%) |
| 5.RA.A.1d | 456 | 135 (29.6%) | 133 (29.2%) |
| 5.RA.A.2 | 414 | 101 (24.4%) | 211 (51.0%) |
| 5.GM.B.4a | 765 | 241 (31.5%) | 635 (83.0%) |
| 5.GM.B.4b | 784 | 151 (19.3%) | 424 (54.1%) |
| 5.GM.A.2 | 545 | 240 (44.0%) | 244 (44.8%) |

*Note.* Level 1 = number (and percentage) of Level 1 retests; Level 2 = number (and percentage) of Level 2 retests.

### 4.4.9 Student Testing Times

During the instructionally embedded window, teachers were expected to provide an adequate amount of time assessing students on concepts taught within instructional cycles. In general, we would expect students to spend an adequate amount of time completing each test form. Further, we also expected students to spend an adequate amount of time assessing on content groups across the window. Kite Student Portal recorded start dates, end dates, and timestamps associated with each test form administered during the instructionally embedded window. Differences in the start times and end times were used to calculate the student-level testing times per test form and aggregate student-level testing times across all content groups.

Table 10 presents the descriptive statistics of time needed (in minutes) for students to complete test forms, by the number of standards assessed on each test form. To minimize skewed results, we included only test forms that were started and completed on the same date. The mean time spent on each test form ranged from 5.97 to 31.10 minutes, while the median time spent on each test form ranged from 3.92 to 21.40 minutes. The minimum time spent on each test form ranged from 0.32 to 9.20 minutes, whereas the maximum time spent on each test form ranged from 269.0 to 508.0 minutes. Overall, we observed a general positive trend in the time spent on each test form and the number of standards included on a test form.

Table 10.

*Distribution Of Time Spent On Each Test Form (in Minutes)*

| No. of standards in content group | *n* | *Min* | *Mdn* | *M* | *Max* | 25Q | 75Q | IQR |
|---|---|---|---|---|---|---|---|---|
| 1 | 2,879 | 0.32 | 3.92 | 5.97 | 269.00 | 2.23 | 7.12 | 4.89 |
| 2 | 3,320 | 0.75 | 5.78 | 9.87 | 357.00 | 3.85 | 9.32 | 5.47 |
| 3 | 5,289 | 0.23 | 9.55 | 13.60 | 508.00 | 6.10 | 15.10 | 9.03 |
| 4 | 2,077 | 1.30 | 11.70 | 17.10 | 384.00 | 7.53 | 18.80 | 11.20 |
| 5 | 1,574 | 2.65 | 16.60 | 24.90 | 413.00 | 10.80 | 26.20 | 15.40 |
| 6 | 606 | 6.12 | 18.90 | 24.00 | 292.00 | 13.20 | 27.00 | 13.80 |
| 7 | 147 | 9.20 | 21.40 | 31.10 | 285.00 | 15.60 | 31.90 | 16.30 |

*Note. N* = Number of tests; Min = Minimum; Max = Maximum; 25Q = lower quartile; 75Q = upper quartile; IQR = interquartile range.

In addition, Table 11 presents the distribution of total time spent testing, in minutes, across all students participating in the 2024–2025 PIE instructionally embedded assessment. The amount of time spent testing ranged from 8.43 to 1,817.00 minutes. The mean time spent testing was 137.00 minutes, and the median time spent testing was 97.00 minutes.

Table 11.

*Distribution of Total Time Spent Testing (in Minutes)*

| *N* | *Min* | *Mdn* | *M* | *Max* | 25Q | 75Q | IQR |
|---|---|---|---|---|---|---|---|
| 1,572 | 8.43 | 97.10 | 137.00 | 1,817.00 | 54.50 | 174.00 | 120.00 |

*Note. N* = Number of students; Min = Minimum; Max = Maximum; 25Q = lower quartile; 75Q = upper quartile; IQR = interquartile range.

## 4.5 User Experience

### 4.5.1 Student Perceptions and Experiences

To learn about students' experiences interacting with the PIE assessments, teachers were given a survey at the end of the instructionally embedded assessment window that included three multiple-choice questions about their perceptions of their students' experiences. Additionally,

students were asked to respond to three open-ended questions at the end of the assessment window.

The teacher survey included questions about teachers' perceptions of students' experience with the PIE assessments. As shown in Table 12, overall, teachers had positive perceptions of student experience with the assessments. The majority of teachers agreed that students had access to all accessible supports necessary to participate in the assessment (87%), the content of the PIE assessment was accessible to students (78%), and student responses to PIE items accurately reflected their knowledge, skills, and understandings (69%).

Table 12.

*Summary of Teachers' Perceptions of Students' Experience With Assessments (*N *= 47)*

| Statement | SD | | D | | A | | SA | | A + SA | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| The content of PIE assessments is accessible to students. | 1 | 2 | 9 | 19 | 26 | 55 | 11 | 23 | 37 | 78 |
| Student responses to PIE items accurately reflect their knowledge, skills, and understandings. | 2 | 4 | 13 | 28 | 28 | 60 | 4 | 9 | 6 | 69 |
| Students have access to all accessibility supports necessary to participate in the assessment. | 3 | 6 | 3 | 6 | 33 | 70 | 8 | 17 | 41 | 87 |

*Note*. SD = *strongly disagree*; D = *disagree*; A = *agree*; SA = *strongly agree.*

Students generally responded positively to PIE, appreciating the lower stress, enjoyable format, and alignment with classroom instruction. They liked the untimed assessments, smaller sections with breaks, and multiple-choice questions. Some, however, found the content frustrating when it was too difficult or unfamiliar. Suggestions included adding engaging features like game breaks, prizes, more colorful visuals, and customizable backgrounds. Students also wanted tools like calculators, read-aloud options, and clearer progress indicators. Feedback on PIE's technology was mixed: while some liked the ease of response entry and result checking, others reported glitches, on-screen keyboard issues, and trouble entering fractions. Most students reported trying their best, using strategies like double-checking and pacing, though a few cited fatigue or bad days as barriers to effort.

## 4.5.2 Teacher Perceptions and Experiences

### *4.5.2.1 Professional Learning and Training*

A total of 70 classroom teachers, test coordinators, and other staff supporting implementation of the pilot study in their school participated in the six e-learning modules. As shown in Table 13, the overall participation rate was 91%. Except for two modules (*Interpreting PIE Assessment Results* [69%] and *Using PIE to Inform Your Instruction* [80%]), the participation rate was 100% or almost 100%.

Table 13.

*Completion Rate of e-Learning Modules*

| Module | Completion rate (*N* = 70) |
|---|---|
| Enriching Your Classroom Instruction with PIE | 100% |
| Creating and Using Content Groups | 100% |
| A Closer Look at PIE Assessments | 99% |
| Administering the PIE Assessment | 99% |
| Interpreting PIE Assessment Results | 69% |
| Using PIE to Inform Your Instruction | 80% |
| Average | 91% |

Table 14 presents the descriptive statistics of time spent (in minutes) on each module. The mean time spent on each module ranged from 285 to 5,732 minutes, while the median varied from 11.6 to 410.9 minutes. The time varied greatly among teachers, with the minimum time spent on each module ranging from 1.0 to 4.6 minutes and the maximum time spent on each module ranging from 27,139 to 181,419 minutes. The large gap between the mean and median time across the modules and the wide distributions of the time spent on each module were because some teachers spent several days on a module (i.e., kept the module open for several days). This may have been because teachers often did not have a consistent block of time to focus on one task and finish it at one time in classrooms.

Table 14.

*Descriptive Statistics of Time Spent, in Minutes, on Each Module* (N *= 70)*

| Module | *M* | *SD* | *Mdn* | *Min* | *Max* |
|---|---|---|---|---|---|
| Enriching Your Classroom Instruction With PIE | 1,055.7 | 4,134.6 | 12.0 | 3.2 | 27,139.0 |
| Creating and Using Content Groups | 2,600.9 | 6,595.4 | 29.1 | 4.6 | 35,588.4 |
| A Closer Look at PIE Assessments | 5,732.0 | 7,914.0 | 410.9 | 2.3 | 30,207.4 |
| Administering the PIE Assessment | 1,425.5 | 4,408.2 | 12.8 | 1.1 | 24,390.6 |
| Interpreting PIE Assessment Results | 3,921.5 | 26,170.6 | 11.9 | 2.5 | 181,419.7 |
| Using PIE to Inform Your Instruction | 285.0 | 829.6 | 11.6 | 1.0 | 4,916.2 |

*Note.* SD= Standard deviation; Med = Median; Min = Minimum; Max = Maximum

Table 15 shows the descriptive statistics of the time spent on each module for teachers who spent fewer than 60 minutes on each module. After excluding teachers who spent more than 60 minutes on each module, the number of teachers in each module ranges from 31 to 59. The mean time ranges from 10.9 (*Using PIE to Inform Your Instruction* module) to 22.8 minutes (*Creating and Using Content Groups* module).

Table 15.

*Descriptive Statistics of Time Spent, in Minutes, on Each Module (Excluding Cases When Time Spent On Each Module Exceeds 60 Minutes)*

| Module | *n* | *M* | *SD* | *Mdn* | *Min* | *Max* |
|---|---|---|---|---|---|---|
| Enriching Your Classroom Instruction With PIE | 59 | 13.2 | 8.9 | 11.4 | 3.2 | 54.6 |
| Creating and Using Content Groups | 47 | 22.8 | 14.8 | 17.4 | 4.6 | 57.9 |
| A Closer Look at PIE Assessments | 31 | 15.3 | 11.9 | 13.1 | 2.3 | 57.1 |
| Administering the PIE Assessment | 55 | 11.6 | 8.2 | 9.7 | 1.1 | 36.1 |
| Interpreting PIE Assessment Results | 42 | 13.6 | 8.6 | 11.1 | 2.5 | 39.2 |
| Using PIE to Inform Your Instruction | 46 | 10.9 | 8.1 | 10.2 | 1.0 | 43.6 |

*Note. N* = Number of teachers; SD= Standard deviation; Med = Median; Min = Minimum; Max = Maximum

At the end of each module, teachers could opt to take a survey about the training module. As shown in Table 16, among the 70 teachers who participated in the training, six (9%) to 14 (20%) teachers chose to take the survey.

Table 16.

*Number of Teachers Participating in Post–e-Learning Moodle Training Survey*

| Module | *n* | Response rate (%) |
|---|---|---|
| A Closer Look at PIE Assessments | 11 | 16 |
| Administering PIE Assessments | 8 | 11 |
| Creating and Using Content Groups | 8 | 11 |
| Enriching Your Classroom Instruction with PIE | 14 | 20 |
| Interpreting PIE Assessment Results | 6 | 9 |
| Using PIE to Inform Your Instruction | 6 | 9 |

*Note.* Response rate indicates the percentage of teacher participants for each posttraining survey among the total number of teachers who participated in the e-learning Moodle training.

Table 17 presents the results of the teacher-survey questions about teachers' perceptions of the PIE training. Overall, teachers had positive perceptions of the training. All teachers but one (46 of 47 teachers) agreed that the PIE training modules and resources/job aids helped them know how to create content groups for assessment and instruction. Eighty-three percent agreed that the PIE training modules and resources/job aids provide them with the knowledge and skills to implement cycles of assessment and instruction. In addition, 83% said they found the PIE resources helpful in understanding the PIE learning pathway content. Similarly, 77% agreed that the training modules and resources/job aids helped them understand how to interpret and use data to inform instruction.

Table 17.

*Summary of Teacher Responses About Teachers' Perceptions of the Training (*N *= 47)*

| Statement | SD | | D | | A | | SA | | A + SA | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| The PIE training modules and resources/job aids helped me know how to create content groups for assessment and instruction. | 0 | 0 | 1 | 2 | 29 | 62 | 17 | 36 | 46 | 98 |
| The PIE training modules and resources/job aids provided me with the knowledge and skills to implement cycles of assessment and instruction. | 0 | 0 | 8 | 17 | 31 | 66 | 8 | 17 | 39 | 83 |
| PIE resources helped me understand the PIE learning pathway content. | 0 | 0 | 8 | 17 | 31 | 66 | 8 | 17 | 39 | 83 |
| The PIE training modules and resources/job aids helped me understand how to interpret and use data to inform instruction. | 1 | 2 | 10 | 21 | 27 | 57 | 9 | 19 | 36 | 76 |

*Note*. SD = *strongly disagree*; D = *disagree*; A = *agree*; SA = *strongly agree*.

*4.5.2.2 Creating PIE Content Groups*

Table 18 shows the results of the teacher survey questions about creating PIE content groups. Among 47 teachers, 85% responded that they were able to use the learning pathways to create content groups for assessment. The same percentage agreed that the PIE content groups management page in Kite Educator Portal was easy to navigate.

Table 18.

*Summary of Teacher Responses About Creating PIE Content Groups (*N *= 47)*

| Statement | SD | | D | | A | | SA | | A + SA | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| I was able to use the learning pathways to create content groups for assessment. | 0 | 0 | 7 | 15 | 25 | 53 | 15 | 32 | 40 | 85 |
| The PIE content groups management page in Kite Educator Portal was easy to navigate. | 1 | 2 | 8 | 17 | 22 | 47 | 16 | 34 | 38 | 81 |

*Note*. SD = *strongly disagree*; D = *disagree*; A = *agree*; SA = *strongly agree*.

In their responses to open-ended survey questions, teachers suggested changes to the content groups management page in Kite Educator Portal. Teachers wanted to be able to set up content groups for all classes at the same time. They also wanted to add more learning targets and add or alter standards after the group had been created (but before testing began). Further, teachers wanted to be able to order the content groups on their own, rather than relying on the default alphabetical order. Teachers also requested the presence of content group suggestions in the portal and the ability to compare and receive feedback on their scope and sequence vs. the content groupings.

*4.5.2.3 Administration Timing and Process*

Table 19 shows the results of teacher surveys about administration timing and process. Among 47 teachers, 79% reported that they could administer assessments at instructionally-relevant points in time. Regarding the length of the assessment window, 56% agreed that the length was sufficient for completing the instruction and assessment cycle for each of their planned content groups, while 44% disagreed.

Table 19.

*Summary of Teacher Responses About Administration Timing & Process (*N *= 47)*

| Statement | SD | | D | | A | | SA | | A + SA | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| I was able to administer assessments at instructionally-relevant points in time (i.e., after instruction on the assessed content had occurred). | 3 | 6 | 7 | 15 | 25 | 53 | 12 | 26 | 37 | 79 |
| The length of the assessment window was sufficient for completing the instruction and assessment cycle for each of my planned content groups. | 10 | 21 | 11 | 23 | 20 | 43 | 6 | 13 | 26 | 56 |
| The administration process became easier over the course of the window. | 0 | 0 | 5 | 11 | 23 | 49 | 19 | 40 | 42 | 89 |

*Note*. SD = *strongly disagree*; D = *disagree*; A = *agree*; SA = *strongly agree*.

In their open-ended survey responses, teachers generally provided positive feedback on administering PIE assessments, with several highlighting the useful reporting dashboard with the inclusion of mastery symbols, and the helpfulness of the learning pathways. They appreciated the assessment design, noting the mix of question types and adaptive sections. Real-time reporting was also valued for supporting teaching. However, some issues were noted, including time constraints (i.e., some were unable to cover all content within the testing window, especially due to unforeseen disruptions), PIE's interface and wait times between tests, the desire to see the actual test questions, and other technological issues, particularly with inputting content groups and student responses, were also noted.

### 4.5.3 Student Opportunity to Learn

Table 20 presents the results of the teacher survey questions about students' opportunity to learn and instruction. Eighty-three percent of teachers agreed that students had the opportunity to learn content aligned with the PIE learning pathways. Regarding instruction, 62% concurred that PIE assessment items were aligned with instruction, and 77% reported being able to use instructional strategies to reach a continuum of learners at various places in the learning pathways.

Table 20.

*Summary of Teacher Responses About Students' Opportunity to Learn (*N *= 47)*

| Statement | SD | | D | | A | | SA | | A + SA | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| Students had the opportunity to learn content aligned with PIE learning pathways. | 0 | 0 | 8 | 17 | 27 | 57 | 12 | 26 | 39 | 83 |
| PIE assessment items are aligned with instruction. | 2 | 4 | 16 | 34 | 21 | 45 | 8 | 17 | 29 | 62 |

*Note*. SD = *strongly disagree*; D = *disagree*; A = *agree*; SA = *strongly agree*.

# Scoring and Reporting

## 5.1 Instructionally Embedded Scoring and Reporting

The statistical model that was used to establish the probability of mastery for each pathway level for PIE assessments during the instructionally embedded assessment window was the HDCM. The HDCM is a general latent-class model that uses a log-linear modeling framework to produce probabilities of class membership for multiple measured attributes, while also restricting the latent classes that are permitted (Templin & Bradshaw, 2014). Student mastery statuses for each pathway level are inferred using a Bayesian estimation method.

### 5.1.1 Model Specification

Each learning pathway was calibrated separately using separate diagnostic classification models (DCMs). Each pathway level within a learning pathway was defined as an attribute, and each model was defined to measure three attributes (i.e., Level 1, Level 2, Level 3). In a DCM framework, a probability of membership in each permitted latent class represented in learning pathway is produced on a continuous scale ranging from 0 to 1. In addition, a marginal probability of mastery for each pathway level in a learning pathway is produced using the class membership probabilities and reported on a continuous scale ranging from 0 to 1. Students were then designated as a master or nonmaster for each pathway level. Students summarized by a marginal posterior probability of mastery that is greater than 0.7 for a particular pathway level were determined as a master of that pathway level.

The general mathematical representation for DCMs is defined as:

$$P(Y_j = y_j) = \sum_{c=1}^{C} v_c \prod_{i=1}^{I} \pi_{ic}^{y_{ji}} (1 - \pi_{ic})^{1 - y_{ji}} \quad (8)$$

The term $\pi_{ic}$ denotes the conditional probability of a student in class $c$ providing a correct response to item $i$. The term $y_{ji}$ denotes the observed response of student $j$ to item $i$. Last, the term $v_c$ denotes the base-rate parameter that describes the proportion of the student population that are members of class $c$.

*5.1.1.1 Log-linear Cognitive Diagnosis Model*

A variety of model parameterizations can be defined for $\pi_{ic}$ to reflect different assumptions about the DCM measurement model (i.e., relationships between the items and the measured attributes). The log-linear cognitive diagnosis model (LCDM; Henson et al., 2009) is a general DCM that mathematically defines the conditional probability of a correct response for a student in class *c* on item *i* using a generalized linear model with a logit-link function. The assessment items administered during the pilot study were designed to measure a single pathway level within a single LP.

For an arbitrary item that measures attribute *a*, the measurement model that describes the probability of a correct response to item *j* for a student in class *c* is defined as:

$$P\left(Y_{ji} = 1 \mid \alpha_c\right) = \frac{\exp\left(\lambda_{i,0} + \lambda_{i,(a)}\alpha_{ca}\right)}{\left[1 + \exp\left(\lambda_{i,0} + \lambda_{i,(a)}\alpha_{ca}\right)\right]} \quad (9)$$

In the above expression, $\alpha_{ca}$ is a binary indicator denoting whether a student in class *c* is a master (e.g., $\alpha_{ca} = 1$) or nonmaster ($\alpha_{ca} = 0$) of attribute *a*. The intercept parameter $\lambda_{i,0}$ denotes the log-odds of a correct response on item *i* for students who have not mastered attribute *a*. The main-effect parameter $\lambda_{i,1,(a)}$ denotes the change in the log-odds of a correct response on item *i* that results from mastering attribute *a*. We expect students who have mastered the measured attribute to answer an item correctly at a high rate compared to students who have not mastered the measured attribute. Therefore, $\lambda_{i,1,(a)}$ was constrained to be a nonnegative value, ensuring monotonicity of the mastery classes.

For graded items (e.g., 2-point items), the LCDM measurement model can be extended to describe the probability of observing a response greater than category *k* to item *j* for a student in class *c*:

$$P\left(Y_{ji} \geq k \mid \alpha_c\right) = \frac{\exp\left(\lambda_{k;\,i,0} + \lambda_{i,(a)}\alpha_{ca}\right)}{\left[1 + \exp\left(\lambda_{k;\,i,0} + \lambda_{i,(a)}\alpha_{ca}\right)\right]}, k = 1, \ldots, K_i - 1, \quad (10)$$

where $K_i$ denotes the number of item categories. Graded items are described by $K_i$ parameters, which includes a single main-effect parameter $\lambda_{i,1,(a)}$ and $K_i - 1$ intercept parameters $\lambda_{k;\,i,0}$. For graded items, the intercept parameters represent category thresholds between adjacent categories k-1 and k. The probability of providing a response to category k can be expressed as

$$P\left(Y_{ji} = k \mid \alpha_c\right) = P\left(Y_{ji} \geq k \mid \alpha_c\right) - P\left(Y_{ji} \geq k + 1 \mid \alpha_c\right), \quad (11)$$

where $P\left(Y_{ji} \geq 0 \mid \alpha_c\right) = 1$ and $P\left(Y_{ji} \geq K \mid \alpha_c\right) = 0$ are fixed to ensure identifiability of the model.

*5.1.1.2 Hierarchical Diagnostic Classification Scoring Model (HDCM)*

Because the models were calibrated at the learning pathway level, eight possible mastery classes exist within an LCDM framework. However, the design of the PIE assessments assumes for each learning pathway that (a) mastery of a Level 1 skill is a precursor to mastering a Level 2 skill, and (b) mastery of a Level 2 skill is a precursor to mastering a Level 3 skill.

This assumption implies a hierarchical ordering of the pathway levels in each learning pathway (e.g., von Davier & Haberman, 2014). Therefore, a hierarchical DCM parameterization was defined for each of the 25 PIE learning pathways. The HDCM excludes attributes profiles that are not aligned with the defined learning pathway hierarchy. Thus, the HDCM reflects a more parsimonious model compared to the LCDM by restricting base-rate parameters associated with impermissible attribute profiles to zero. Four possible mastery classes exist with the HDCM framework for each of the 25 learning pathways in PIE. Therefore, three base-rate parameters were estimated for each model, and the fourth base-rate parameter was deterministically calculated according to the three freely estimated base-rate parameters.

### 5.1.2 Model Estimation

The DCMs were estimated using a Bayesian estimation approach. A Bayesian estimation approach was selected over traditional maximum-likelihood-estimation approaches because a Bayesian approach offers more robust methods for evaluating model fit (Thompson, 2023). Twenty-five learning pathways were assessed in the pilot study, resulting in 25 total models. Each model was estimated using a Markov Chain Monte Carlo procedure using the software package *rstan* (Stan Development Team, 2024), an interface to the *Stan* probabilistic programming language (Carpenter et al., 2017). The models were estimated with four chains, using 4,000 iterations for each chain. The first 2,000 iterations were used as warm-up in each chain and removed, resulting in 8,000 total iterations producing the posterior distributions.

The Bayesian estimation approach described here defines the same priors for each model for each learning pathway. The intercept parameters are unbounded, and a normal distribution with a mean of 0 and a variance of 2 is used as a prior. The main-effect parameters are constrained to be positive to ensure monotonicity holds, and a lognormal distribution with a mean of 0 and a variance of 1 is used as a prior.

After estimating each model, we checked that the models had converged (i.e., whether the sampling algorithm sufficiently explored the posterior parameter space) by assessing $\hat{R}$ and effective sample-size evaluation criteria (see Vehtari et al., 2021). When determining convergence, we checked that all $\hat{R}$ values were below 1.01 and that all effective sample sizes were greater than 400, as described in Vehtari et al. (2021). Following the model evaluation procedure, we calculated parameter estimates using the mean of the posterior distribution and then used those parameters to estimate marginal student probabilities of mastery for each measured pathway level in the PIE assessment.

### 5.1.3 Estimation of Student Mastery Probabilities

Student response data from the PIE assessments were used to estimate posterior probabilities of mastery for assessed pathway levels for students using the calibrated HDCM models. For the PIE assessment scoring procedures, expected a posterior (EAP) estimates were used to report student mastery probabilities. To calculate EAP estimates, we first calculated the probability that a student belonged to each of the four attribute classes. For student *j*, the probability of belonging to class *c* is defined as:

$$\hat{\alpha}_{jc} = \frac{\nu_c \prod_{i=1}^{I} [\pi_{ic}^{y_{ji}} (1-\pi_{ic})^{1-y_{ji}}]}{\sum_{c=1}^{4} \nu_c \prod_{i=1}^{I} [\pi_{ic}^{y_{ji}} (1-\pi_{ic})^{1-y_{ji}}]} \quad (12)$$

For student j and pathway level a, the EAP estimate was calculated using the class membership probabilities as:

$$\hat{p}_{ja} = \sum_{c=1}^{4} \hat{\alpha}_{jc} \alpha_c. \quad (13)$$

The threshold used to determine mastery classifications for each pathway level was set to 0.7. A threshold of 0.7 was used to balance the risk of Type 1 and Type 2 errors resulting from the classification-based scoring procedure.

## 5.1.4 Model Evaluation

### *5.1.4.1 Absolute Model Fit*

Posterior predictive model checks were used to assess absolute model fit using a simulation-based approach (e.g., Thompson, 2019). Using the values obtained from the Bayesian estimation approach at each iteration of the sampling algorithm, 8,000 replicated data sets were generated for each model and compared to the original data. For each iteration of the chain,

1. Each student was randomly assigned to one of four attribute classes according to estimated class membership probabilities, and
2. A response to each administered item for each student was simulated according to their simulated attribute class in (1) and the estimated conditional probability of correctly answering each item, calculated using the estimated parameters.

This procedure was conducted for each iteration of the chain. Summaries of the replicated data sets were then used to assess how closely the estimated model captured patterns in the observed data.

At the model level, we calculated the number of students expected by the model at each raw-score point (i.e., the number of points scored across all items) in the raw-score distribution. Model misfit is assessed by calculating two types of $\chi^2$ discrepancy measure (see Béguin & Glas, 2001). First, a $\chi^2$ discrepancy measure is calculated for the observed data set as

$$\chi^2_{obs} = \sum_{s=0}^{S} \frac{(n_s - E[n_s])^2}{E[n_s]}, \quad (14)$$

where *s* denotes the score point, $n_s$ denotes the number of students that achieved score point *s*, and $E[n_s]$ denotes the number of students expected to perform at each score point according to the model. $E[n_s]$ is calculated as the average number of students expected to perform at each score point across all replicated data sets. Second, a $\chi^2$ discrepancy measure is calculated for each replicated data set as

$$\chi^2_{rep} = \sum_{s=0}^{S} \frac{(n_s - E[n_s])^2}{E[n_s]}. \quad (15)$$

In the discrepancy measure calculated for a particular replicated data set, $n_s$ denotes the number of students that achieved score point *s* in that replicated data set. Finally, a posterior

predictive *p* (ppp) value is calculated by comparing $\chi^2_{obs}$ to a reference distribution formed from the 8,000 $\chi^2_{rep}$ statistics,

$$ppp = P(\chi^2_{rep} > \chi^2_{obs} \mid n_s). \quad (16)$$

The ppp value quantifies how closely the estimated model captures patterns in the observed data. Low ppp values can illustrate that a model poorly fits the data. High ppp values can illustrate potential overfitting of the model to the data.

In Table 21, we report the ppp values calculated for each learning pathway using the calibrated HDCMs. Because the complete set of models across the 25 learning pathways was assessed simultaneously, we used a Holm correction (Holm, 1979) to calculate adjusted ppp values control for familywise error rates. We used a threshold value of .05 for adjusted ppp values to determine model misfit to maintain consistency with existing operational DCM assessment systems (e.g., DLM Consortium, 2022). Calibrated models with an adjusted ppp value less than .05 was flagged as a poorly fitting model. Among the 25 calibrated HDCMs, 20 (80%) were flagged as demonstrating adequate absolute model fit. Among the five learning pathways at which the HDCMs demonstrated less than adequate model fit, four of the flagged learning pathways were from the number sense and operations in fractions domain and one flagged learning pathway was from the data and statistics domain.

Table 21.

*Summary of Absolute Fit for Calibrated Hierarchical Diagnostic Classification Scoring Models*

| Learning pathway | Posterior predictive *p* value | Adjusted *p* value |
|---|---|---|
| PIE.5.NF.A.1 | 0.000 | 0.000 |
| PIE.5.NF.A.2 | 0.187 | 1.000 |
| PIE.5.NF.A.3 | 0.531 | 1.000 |
| PIE.5.NF.B.4 | 0.018 | 0.314 |
| PIE.5.NF.B.5a | 0.604 | 1.000 |
| PIE.5.NF.B.5b | 0.001 | 0.016 |
| PIE.5.NF.B.5c | 0.000 | 0.006 |
| PIE.5.NF.B.5d | 0.885 | 1.000 |
| PIE.5.NF.B.6 | 0.477 | 1.000 |
| PIE.5.NF.B.7a | 0.005 | 0.095 |
| PIE.5.NF.B.7b | 0.394 | 1.000 |
| PIE.5.NF.B.7c | 0.011 | 0.196 |
| PIE.5.NF.B.8a | 0.581 | 1.000 |
| PIE.5.NF.B.8b | 0.000 | 0.000 |
| PIE.5.RA.A.1a | 0.897 | 1.000 |
| PIE.5.RA.A.1b | 0.652 | 1.000 |
| PIE.5.RA.A.1c | 0.941 | 1.000 |
| PIE.5.RA.A.1d | 0.272 | 1.000 |
| PIE.5.RA.A.2 | 0.881 | 1.000 |
| PIE.5.RA.C.5 | 0.399 | 1.000 |
| PIE.5.GM.A.2 | 0.009 | 0.164 |
| PIE.5.GM.B.4a | 0.288 | 1.000 |
| PIE.5.GM.B.4b | 0.338 | 1.000 |
| PIE.5.GM.C.6a | 0.418 | 1.000 |
| PIE.5.DS.A.2 | 0.002 | 0.032 |

#### *5.1.4.2 Classification Accuracy and Classification Consistency*

Reliability measures provide an important quantification for the precision of scores produced within the PIE assessment. In DCM, reliability measures reflect the level of agreement between estimated and true mastery classifications. We report two types of reliability for DCMs: classification accuracy and classification consistency.

*Classification accuracy* refers to how correctly students are classified into their true latent classes according to their observed responses. As defined by Johnson and Sinharay (2018), it is the probability that a student is placed in the correct latent class, given the model parameter estimates. In DCMs, this is a key indicator of model fit. It reflects how accurate the model's classification decisions are for each attribute, typically summarized using a 2 × 2 table comparing true and estimated mastery statuses. For PIE assessments, classification accuracy is evaluated at the pathway level and helps determine how confidently we can interpret students' skill classifications.

As described by Johnson and Sinharay (2018), classification accuracy can be estimated for each pathway level as:

$$\widehat{P_A} = \frac{1}{N}\sum_{n=1}^{N} \tilde{\alpha} \cdot P(\alpha = 1 \mid X = x_n) + \frac{1}{N}\sum_{n=1}^{N}(1 - \tilde{\alpha}) \cdot P(\alpha = 0 \mid X = x_n), \qquad (17)$$

where *N* is the number of students, $\tilde{\alpha}$ is the mastery status implied by the model, and $P(\alpha = 1 \mid \boldsymbol{X} = \boldsymbol{x_n})$ and $P(\alpha = 0 \mid \boldsymbol{X} = \boldsymbol{x_n})$ are the probabilities that a pathway level was mastered and not mastered, respectively, under the estimated model. We used Johnson and Sinharay's (2018) guidelines for interpreting classification accuracy values: ≥ .99 is *excellent*, .95–.98 is *very good*, .89–.94 is *good*, .83–.88 is *fair*, .55–.82 is *poor*, and < .55 is *weak*.

In total, 49 (65.3%) pathway levels demonstrated at least *fair* classification accuracy. Table 22 presents the number of models within each pathway level that demonstrated each category of classification accuracy. The results show that less than half of level 1 skills showed at least *fair* classification accuracy, with rates increasing for higher pathway levels.

Table 22.

*Number of Learning Pathways by Estimated Classification Accuracy Category and Pathway Level*

| Pathway level | *Weak* | *Poor* | *Fair* | *Good* | *Very good* | *Excellent* |
|---|---|---|---|---|---|---|
| | .00–.54 | .55–.82 | .83–.88 | .89–.94 | .95–.98 | .99–1.00 |
| Level 1 | 0 | 14 | 5 | 4 | 2 | 0 |
| Level 2 | 0 | 9 | 11 | 5 | 0 | 0 |
| Level 3 | 0 | 3 | 14 | 8 | 0 | 0 |

*Classification consistency* is a measure of how consistently a student would be placed into the same latent class if the assessment were administered again, assuming the model fit is adequate. Rather than using a repeated testing approach, classification consistency can be calculated using the estimated model parameters from a single administration. Johnson and Sinharay (2018) defined classification consistency as the probability that a student would receive the same classification across repeated test administrations. We used Johnson and Sinharay's guidelines for interpreting classification consistency values: 1.0 is *excellent*, .99 is *very good*, .86–.98 is *good*, .70–.85 is *fair*, and < .70 is *poor*.

In total, 66 (88.0%) pathway levels demonstrated at least *fair* classification consistency. Table 23 presents the number of models within each pathway level that demonstrated each category of classification consistency. The results show a similar pattern across all pathway levels, with no level consistently exhibiting higher or lower consistency than the others.

Table 23.

*Number of Learning Pathways by Estimated Classification Consistency Category and Pathway Level*

| Pathway level | *Poor* | *Fair* | *Good* | *Very good* | *Excellent* |
|---|---|---|---|---|---|
| | .00–.69 | .70–.85 | .86–.98 | .99 | 1.00 |
| Level 1 | 3 | 5 | 6 | 2 | 9 |
| Level 2 | 4 | 8 | 5 | 0 | 8 |
| Level 3 | 2 | 8 | 7 | 0 | 8 |

## 5.1.5 Calibrated Parameters

For PIE assessments, the item parameters are the conditional probability of nonmasters scoring a value of 1 or greater (defined as the inverse logit of $\lambda_{1;\,i,0}$), the conditional probability of nonmasters scoring a value of 2 (defined as the inverse logit of $\lambda_{2;\,i,0}$), the conditional probability of masters scoring a value of 1 or greater (defined as the inverse logit of ($\lambda_{1;\,i,0} + \lambda_{i,1,(a)}$)), and the conditional probability of masters scoring a value of 2 (defined inverse logit of ($\lambda_{2;\,i,0} + \lambda_{i,1,(a)}$)). Note that for binary items, the conditional probability of scoring a value of 2 is equal to zero. Using Equations 10-11, the conditional probability of scoring a particular value can be computed using the item parameters. Then, an average score for masters and nonmasters can be calculated for the items. In addition to the item parameters, the DCMs also include a structural parameter, which defines the base rates of class membership for each standard. The item and structural parameters were obtained by calibrating HDCMs for each standard using response data from the 2024 PIE field test (see ATLAS, 2025a). A summary of the parameters used to score-response data from the 2024–2025 instructionally embedded assessments is provided in the next sections.

### *5.1.5.1 Response Probabilities for Masters*

When items measuring a given pathway level function as intended, students who have mastered that level should exhibit a high average score on the items measuring that level. However, if items that produce low average scores for masters may indicate that the items do not effectively differentiate between mastery levels or are misaligned with the intended cognitive construct, which may result in masters erroneously being labeled as nonmasters. Such misclassification can significantly compromise the validity of the assessment, leading to inaccurate conclusions about student learning. Most critically, it can result in misguided instructional decisions (i.e., unnecessary retesting) that affect student progress during the instructionally embedded window.

Figure 26 presents the average score of masters on 1-point items. Figure 27 presents the average score of masters on 2-point items. The maximum point of uncertainty corresponds to an average score of 0.5 times the maximum number of possible points for an item. For items functioning as expected, masters should have an average score above 0.5 and 1.0 for 1-point and 2-points items, respectively. In total, 388 (90.9%) items exhibited an average score above the defined threshold, while 39 (9.1%) items exhibited an average score below the defined

threshold. These findings suggest that a vast majority of the assessment items functioned as expected for students who mastered the given pathway level.

Figure 26.

*Average Score of Masters for Math 1-Point Items*

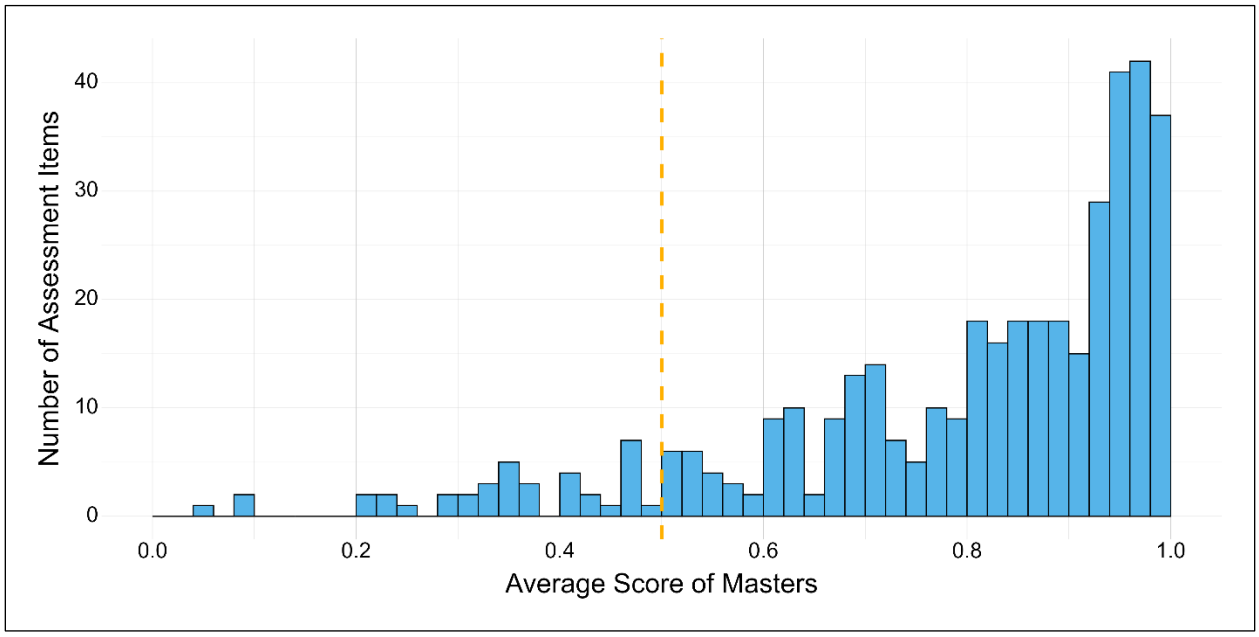
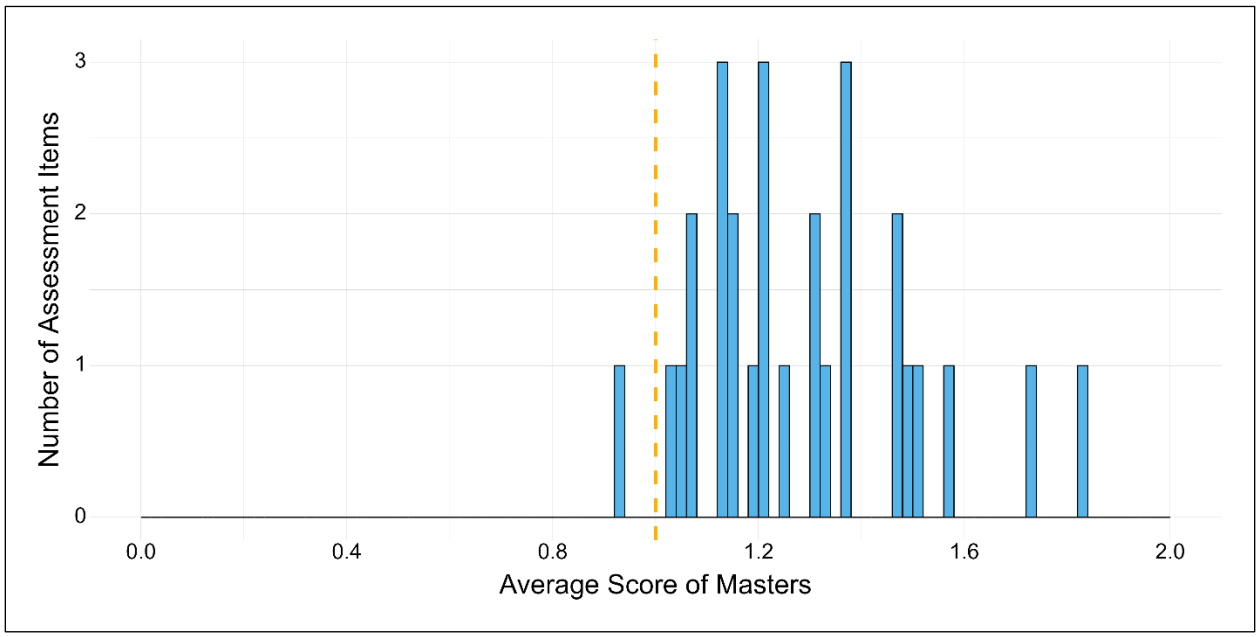


Figure 27.

*Average Score of Masters for Math 2-Point Items*




*5.1.5.2 Response Probabilities for Nonmasters*

When items measuring a given pathway level function as intended, students who have not mastered that level should exhibit low average scores on those items. These lower expected values indicate that nonmasters are appropriately selecting response options that reflect their current level of understanding. However, items that yield unexpectedly high average scores for nonmasters may signal that the items do not effectively distinguish between mastery levels or are misaligned with the intended cognitive construct. In such cases, nonmasters may be erroneously classified as masters. Such misclassifications significantly compromise the validity of the assessment, leading to overestimations of student proficiency and resulting in missed instructional opportunities to address foundational gaps. These inaccuracies can hinder students' academic progress in the instructionally embedded window by allowing them to advance without the necessary support.

Figure 28 presents the average score of nonmasters on 1-point items. Figure 29 presents the average score of nonmasters on 2-point items. The maximum point of uncertainty corresponds to an average score of 0.5 times the maximum number of possible points for an item. For items functioning as expected, nonmasters should have an average score below 0.5 and 1.0 for 1-point and 2-point items, respectively. In total, 334 (78.2%) items exhibited an average score below the defined threshold, while 93 (21.8%) items exhibited an average score above the defined threshold. These findings suggest that a vast majority of the assessment items functioned as expected for students who had not mastered the given pathway level. However, the number of items functioning as expected for nonmasters was lower than that for masters.

Figure 28.

*Average Score of Nonmasters for 1-Point Items*

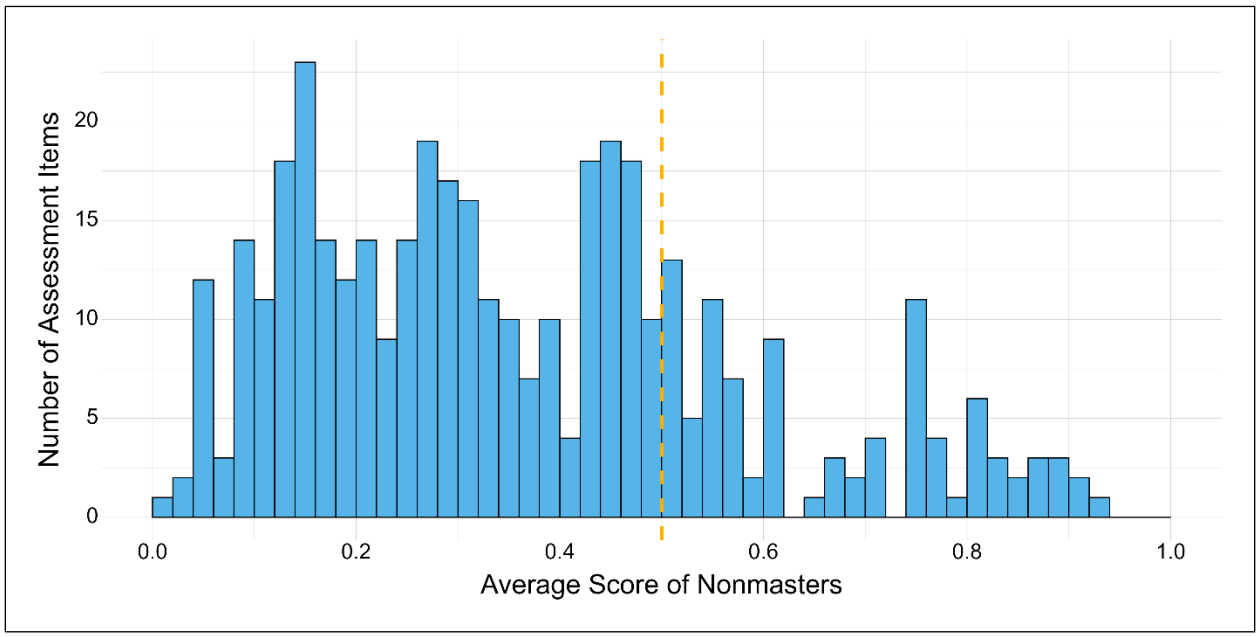
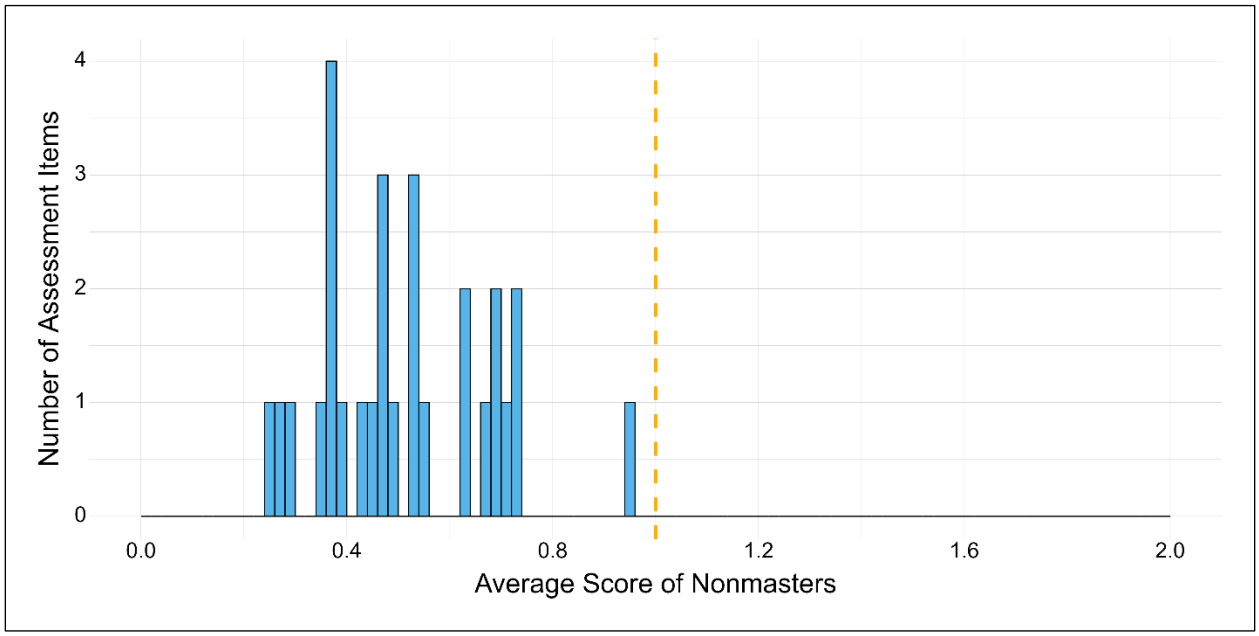



Figure 29.

*Average Score of Nonmasters for 2-Point Items*

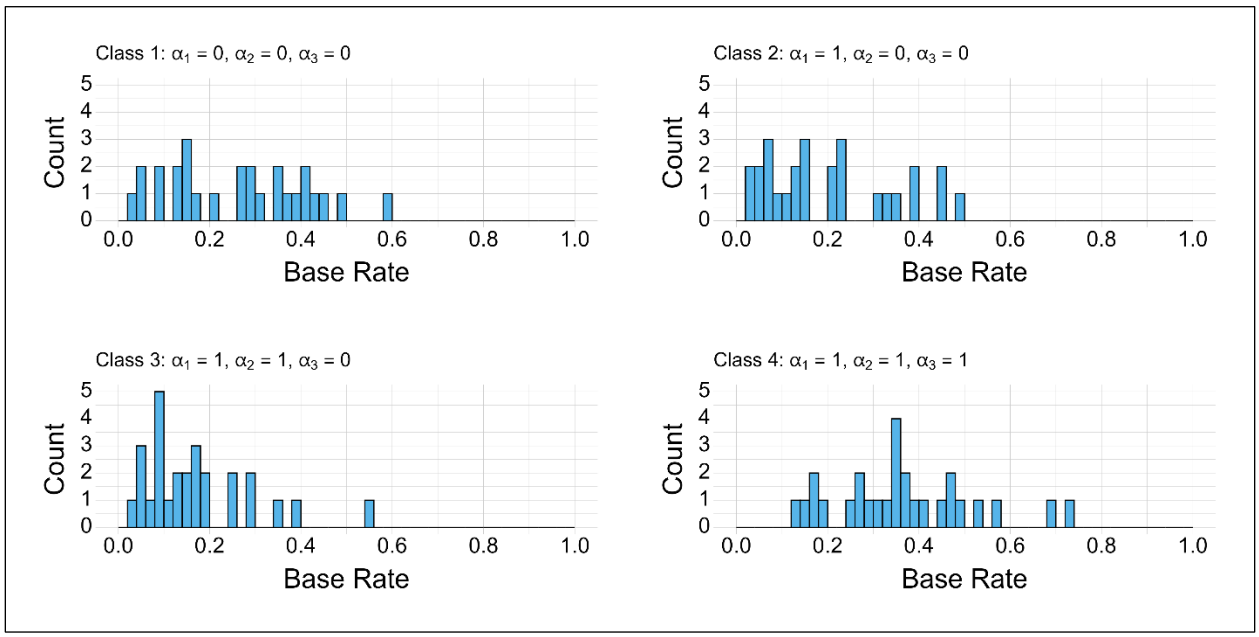


*5.1.5.3 Base-Rate Probability of Class Membership*

The *base rate of class membership* is a structural parameter in DCMs that reflects the estimated proportion of students assigned to each latent class. Figure 30 shows the distribution of the base-rate probabilities for each latent class.

Figure 30.

*Distribution of Base-Rate Probabilities for Standard-Level Models*




## 5.1.6 Reports

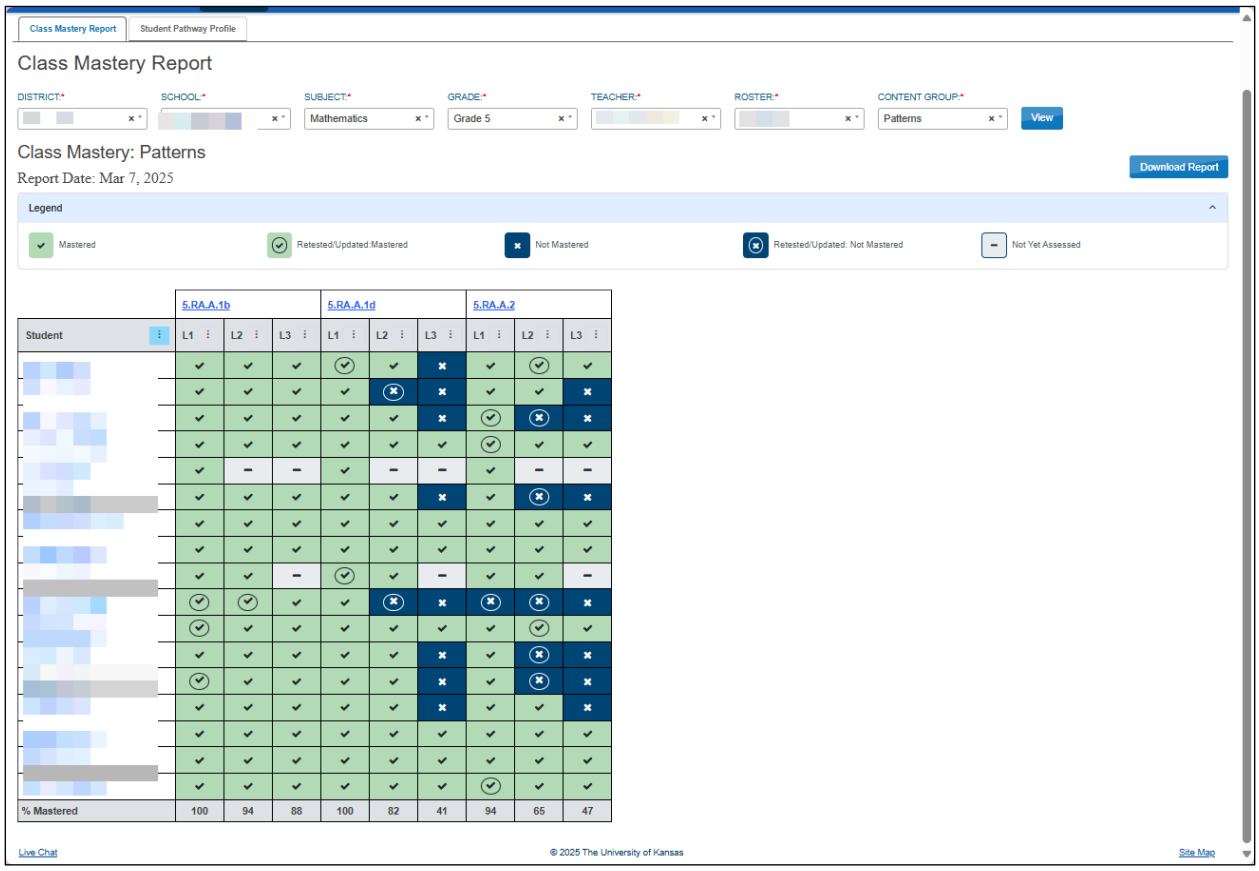
Within the instructionally embedded assessment window, after students completed their assessments, mastery results were automatically available for teachers to review in the PIE Reporting Dashboard in the Kite system. Teachers were encouraged to use this dashboard throughout the instructionally embedded window, as it provided insights to help them identify areas of focus for each student. The dashboard consisted of two types of reports: the Class Mastery Report and the Student Pathway Profile.

In December 2023, PIE staff conducted pre-pilot focus group sessions with educators; we asked them for their input on a prototype version of the reporting dashboard. PIE staff sought educator feedback on the design and features of the dashboard, including the symbols and mastery language used, the information contained in the student profile, and the utility of results presented in the dashboard. Educators gave valuable feedback and suggested improving clarity in mastery language, and layout of the reports. Feedback from these focus group sessions was incorporated into the final reporting dashboard design used during the pilot study.

The Class Mastery Report offers an overview of the entire class for each content group, showing which pathway levels students had mastered for each standard and which ones they had not. Teachers could filter the report grid and download a printable version for further analysis. Additionally, a legend was provided to help teachers understand the meaning of each status.

Figure 31.

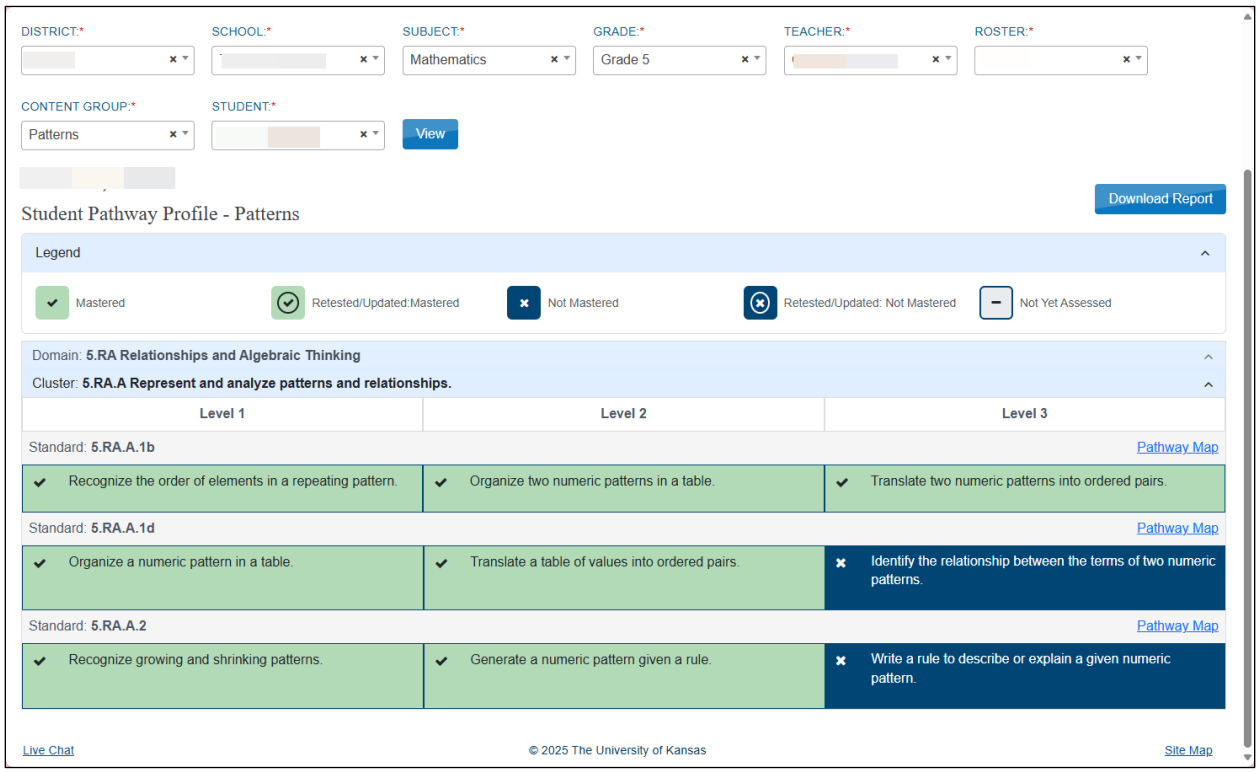
*Sample Class Mastery Report*




The Student Pathway Profile consisted of similar information at an individual student level, offering teachers the ability to download and print them for individual student use. Teachers also could download or view the Pathway Map for each standard, which displayed typical learning progression.

Figure 32.

*Example Student Pathway Profile*




### *5.1.6.1 Educator Feedback*

Table 24 shows the results of teacher surveys on the use of score reports. Overall, teachers found the score reports useful. Among 47 teachers, 72% agreed that the assessment results accurately reflected their perceptions of students' mastery. Similarly, teachers reported that score reports provided instructionally relevant information (74%) and could be used to plan the next steps for instruction (79%). About 90% indicated that the score reports provided one source of evidence of student learning (89%) and that the PIE reporting dashboard in Kite Educator Portal was easy to navigate (87%).

Table 24.

*Summary of Teacher Responses About Use of Score Reports (*N *= 47)*

| Statement | SD | | D | | A | | SA | | A + SA | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| The assessment results are an accurate reflection of my perceptions of students' mastery. | 3 | 6 | 10 | 21 | 31 | 66 | 3 | 6 | 34 | 72 |
| Score reports provide instructionally relevant information. | 2 | 4 | 10 | 21 | 28 | 60 | 7 | 15 | 35 | 74 |
| I can use score reports to plan the next steps for instruction. | 2 | 4 | 8 | 17 | 29 | 62 | 8 | 17 | 37 | 79 |
| Score reports provide me with one source of evidence of student learning. | 2 | 4 | 3 | 6 | 33 | 70 | 9 | 19 | 42 | 89 |
| The PIE reporting dashboard in Kite Educator Portal is easy to navigate. | 3 | 6 | 3 | 6 | 26 | 55 | 15 | 32 | 41 | 87 |

*Note*. SD = *strongly disagree*; D = *disagree*; A = *agree*; SA = *strongly agree*.

Teachers expressed a need for clearer report information, including explanations of checks, colors, and missed items, as well as better insight into whether incorrect answers stemmed from software use or student misunderstanding. They requested more details about question types and depth, with some asking for sample questions. Concerns included test functionality problems, confusion over the 24-hour test wait period, and frustration with the requirement for all students to finish before assigning new assessments. Teachers wanted improved content-grouping tools, such as suggestions; the ability to edit standards, add targets, reorder groups, and compare scope and sequence with standards. While content was praised, it sometimes misaligned with pacing charts. Additional feedback included calls for better module access, a more user-friendly website and training, a longer assessment window, and integration of supporting standards into PIE assessments for use throughout instructional units.

## 5.2 Spring Scoring and Reporting

Scoring and reporting approaches differed between the instructionally embedded and spring windows. In the instructional window, results were generated using a DCM framework. In the spring window, because of the limited item pool per standard, student performance was reported using the number and percentage of points achieved, both overall and by domain. The spring assessment design included four domains totaling 61 points: number sense and operations in fractions (37 points), relationships and algebraic thinking (13 points), geometry and measurement (8 points), and data and statistics (3 points).

### 5.2.1 Reports

After the final assessment, a reporting file was generated for each teacher, which summarized student performance. For every rostered student, the file provided the total points earned and the percentage of possible points achieved, shown both overall and by content domain. The files were distributed to teachers through the secure HTTP site after the close of the assessment window.

## 5.3 Assessment Results

### 5.3.1 Overall Instructionally Embedded Performance

In Table 25, we report a breakdown of the number (and percentage) of students that attained mastery of each pathway level for each content standard. For Level 1 skills, the percentage of students who achieved mastery ranged from approximately 66.1% to 99.4%. For Level 2 skills, the percentage of students who achieved mastery ranged from approximately 48.1% to 97.6%. For Level 3 skills, the percentage of students who achieved mastery ranged from approximately 20.6% to 79.6%. A student must have completed baseline, midway, and end-of-unit assessments for a standard to be included in the performance results.

Table 25.

*Summary of Student Mastery Performance for Level 1, Level 2, and Level 3 Skills*

| Content Standard | Students (*n*) | Level 1 mastery | Level 2 mastery | Level 3 mastery |
|---|---|---|---|---|
| 5.NF.B.7b | 467 | 464 (99.4%) | 404 (86.5%) | 207 (44.3%) |
| 5.NF.B.7c | 499 | 495 (99.2%) | 470 (94.2%) | 397 (79.6%) |
| 5.NF.B.8b | 397 | 367 (92.4%) | 231 (58.2%) | 156 (39.3%) |
| 5.NF.B.5b | 491 | 424 (86.4%) | 378 (77.0%) | 202 (41.1%) |
| 5.DS.A.2 | 389 | 290 (74.6%) | 236 (60.7%) | 168 (43.2%) |
| 5.NF.B.5a | 409 | 396 (96.8%) | 343 (83.9%) | 175 (42.8%) |
| 5.NF.B.5c | 477 | 402 (84.3%) | 294 (61.6%) | 206 (43.2%) |
| 5.NF.B.5d | 623 | 412 (66.1%) | 385 (61.8%) | 209 (33.5%) |
| 5.NF.B.7a | 407 | 400 (98.3%) | 328 (80.6%) | 275 (67.6%) |
| 5.NF.B.8a | 348 | 341 (98.0%) | 240 (69.0%) | 194 (55.7%) |
| 5.NF.A.1 | 1,250 | 1,180 (94.4%) | 1,087 (87.0%) | 861 (68.9%) |
| 5.NF.A.2 | 1,054 | 799 (75.8%) | 713 (67.6%) | 346 (32.8%) |
| 5.NF.A.3 | 977 | 686 (70.2%) | 480 (49.1%) | 219 (22.4%) |
| 5.NF.B.4 | 838 | 623 (74.3%) | 551 (65.8%) | 173 (20.6%) |
| 5.NF.B.6 | 965 | 887 (91.9%) | 565 (58.5%) | 379 (39.3%) |
| 5.RA.C.5 | 333 | 251 (75.4%) | 198 (59.5%) | 108 (32.4%) |
| 5.RA.A.1c | 758 | 685 (90.4%) | 590 (77.8%) | 488 (64.4%) |
| 5.GM.C.6a | 671 | 623 (92.8%) | 466 (69.4%) | 207 (30.8%) |
| 5.RA.A.1a | 619 | 501 (80.9%) | 456 (73.7%) | 206 (33.3%) |
| 5.RA.A.1b | 713 | 704 (98.7%) | 696 (97.6%) | 535 (75.0%) |
| 5.RA.A.1d | 532 | 494 (92.9%) | 426 (80.1%) | 212 (39.8%) |
| 5.RA.A.2 | 529 | 482 (91.1%) | 305 (57.7%) | 222 (42.0%) |
| 5.GM.B.4a | 801 | 697 (87.0%) | 385 (48.1%) | 263 (32.8%) |
| 5.GM.B.4b | 820 | 790 (96.3%) | 545 (66.5%) | 333 (40.6%) |
| 5.GM.A.2 | 560 | 445 (79.5%) | 384 (68.6%) | 207 (37.0%) |

### 5.3.2 Retest Instructionally Embedded Performance

In Table 26 and Table 27, we report a breakdown of the number (and percentage) of students that attained mastery upon retest of Level 1 and Level 2 skills for each learning pathway at midway and end-of-unit assessments, respectively. For Level 1 skills, the percentage of students who achieved mastery upon retest ranged from approximately 19.0% to 74.5%. For Level 2 skills, the percentage of students who achieved mastery upon retest ranged from approximately 7.8% to 77.3%. Similar to the overall performance results, a student must have completed baseline, midway, and end-of-unit assessments for a standard to be included in the retest performance results.

Table 26.

*Summary of Retest Performance for Level 1 Skills*

| Content standard | Students (*n*) | Level 1 mastery after retest |
|---|---|---|
| 5.NF.B.7b | 34 | 25 (73.5%) |
| 5.NF.B.7c | 29 | 17 (58.6%) |
| 5.NF.B.8b | 104 | 63 (60.6%) |
| 5.NF.B.5b | 200 | 106 (53.0%) |
| 5.DS.A.2 | 153 | 29 (19.0%) |
| 5.NF.B.5a | 36 | 16 (44.4%) |
| 5.NF.B.5c | 249 | 139 (55.8%) |
| 5.NF.B.5d | 487 | 176 (36.1%) |
| 5.NF.B.7a | 22 | 14 (63.6%) |
| 5.NF.B.8a | 20 | 9 (45.0%) |
| 5.NF.A.1 | 211 | 116 (55.0%) |
| 5.NF.A.2 | 624 | 191 (30.6%) |
| 5.NF.A.3 | 509 | 204 (40.1%) |
| 5.NF.B.4 | 406 | 122 (30.0%) |
| 5.NF.B.6 | 186 | 94 (50.5%) |
| 5.RA.C.5 | 152 | 47 (30.9%) |
| 5.RA.A.1c | 152 | 47 (30.9%) |
| 5.GM.C.6a | 216 | 137 (63.4%) |
| 5.RA.A.1a | 273 | 85 (31.1%) |
| 5.RA.A.1b | 114 | 89 (78.1%) |
| 5.RA.A.1d | 158 | 100 (63.3%) |
| 5.RA.A.2 | 250 | 106 (42.4%) |
| 5.GM.B.4a | 251 | 110 (43.8%) |
| 5.GM.B.4b | 165 | 123 (74.5%) |
| 5.GM.A.2 | 250 | 106 (42.4%) |

Table 27.

*Summary of Retest Performance for Level 2 Skills*

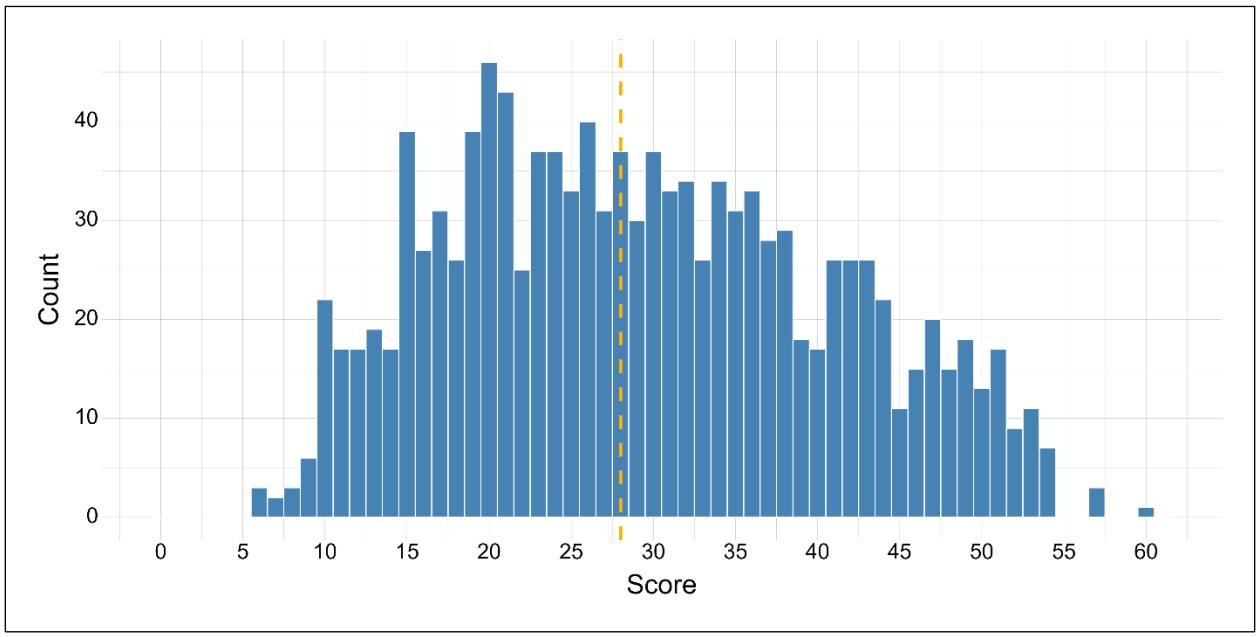
| Content standard | Students (*n*) | Level 2 mastery after retest |
|---|---|---|
| 5.NF.B.7b | 171 | 108 (63.2%) |
| 5.NF.B.7c | 93 | 64 (68.8%) |
| 5.NF.B.8b | 230 | 64 (27.8%) |
| 5.NF.B.5b | 208 | 95 (45.7%) |
| 5.DS.A.2 | 270 | 117 (43.3%) |
| 5.NF.B.5a | 134 | 68 (50.7%) |
| 5.NF.B.5c | 316 | 133 (42.1%) |
| 5.NF.B.5d | 354 | 116 (32.8%) |
| 5.NF.B.7a | 166 | 87 (52.4%) |
| 5.NF.B.8a | 225 | 117 (52.0%) |
| 5.NF.A.1 | 288 | 125 (43.4%) |
| 5.NF.A.2 | 662 | 321 (48.5%) |
| 5.NF.A.3 | 539 | 42 (7.8%) |
| 5.NF.B.4 | 420 | 133 (31.7%) |
| 5.NF.B.6 | 565 | 165 (29.2%) |
| 5.RA.C.5 | 179 | 44 (24.6%) |
| 5.RA.A.1c | 330 | 162 (49.1%) |
| 5.GM.C.6a | 365 | 160 (43.8%) |
| 5.RA.A.1a | 306 | 143 (46.7%) |
| 5.RA.A.1b | 75 | 58 (77.3%) |
| 5.RA.A.1d | 167 | 61 (36.5%) |
| 5.RA.A.2 | 296 | 72 (24.3%) |
| 5.GM.B.4a | 671 | 255 (38.0%) |
| 5.GM.B.4b | 450 | 175 (38.9%) |
| 5.GM.A.2 | 254 | 78 (30.7%) |

### 5.3.3 Spring Performance

Figure 33 presents the distribution of students' overall scores on the assessment, with a maximum possible score of 61 points. The median score was 28 points. The distribution shows a broad spread of performance, with 80% of the students scoring between 15 and 46 points. This pattern suggests notable variability in student performance, with many students clustering around the middle range but fewer achieving scores at the upper end of the scale.

Figure 33.

*Distribution of Overall Scores (Total Possible = 61 Points)*




A more detailed perspective can be obtained by examining the score distributions within individual content domains. Figure 34, Figure 35, Figure 36, and Figure 37 illustrate these distributions for the number sense and operations in fractions, relationships and algebraic thinking, geometry and measurement, and data and statistics domains, respectively. For the number sense and operations in fractions domain, 215 (18.1%) students achieved 70% or more of the total possible points. For the relationships and algebraic thinking domain, 311 (26.2%) students achieved 70% or more of the total possible points. For the geometry and measurement domain, 131 (11.0%) students achieved 70% or more of the total possible points. For the data and statistics domain, 116 (9.8%) students achieved 70% or more of the total possible points. These results indicate that student performance varied notably across content domains, with the highest proficiency observed in the relationships and algebraic thinking domain and the lowest in the data and statistics domain.

Figure 34.

*Distribution of Scores in Number Sense and Operations in Fractions Domain (Total Possible = 37 Points)*

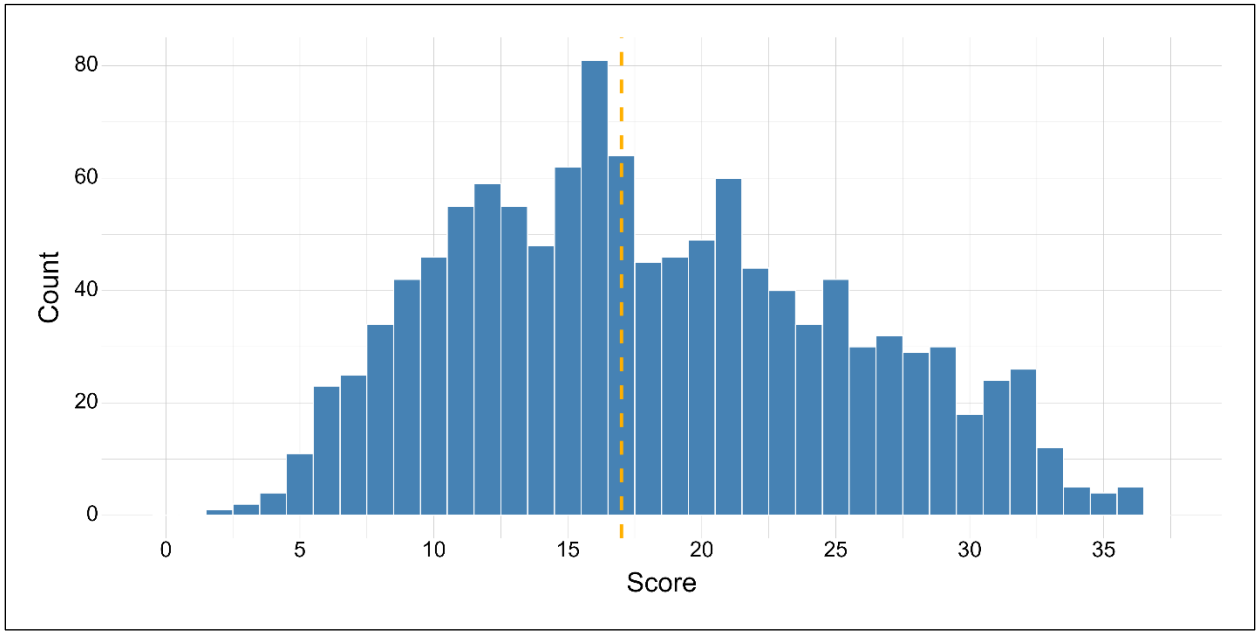



Figure 35.

*Distribution of Scores in Relationships and Algebraic Thinking Domain (Total Possible = 13 Points)*

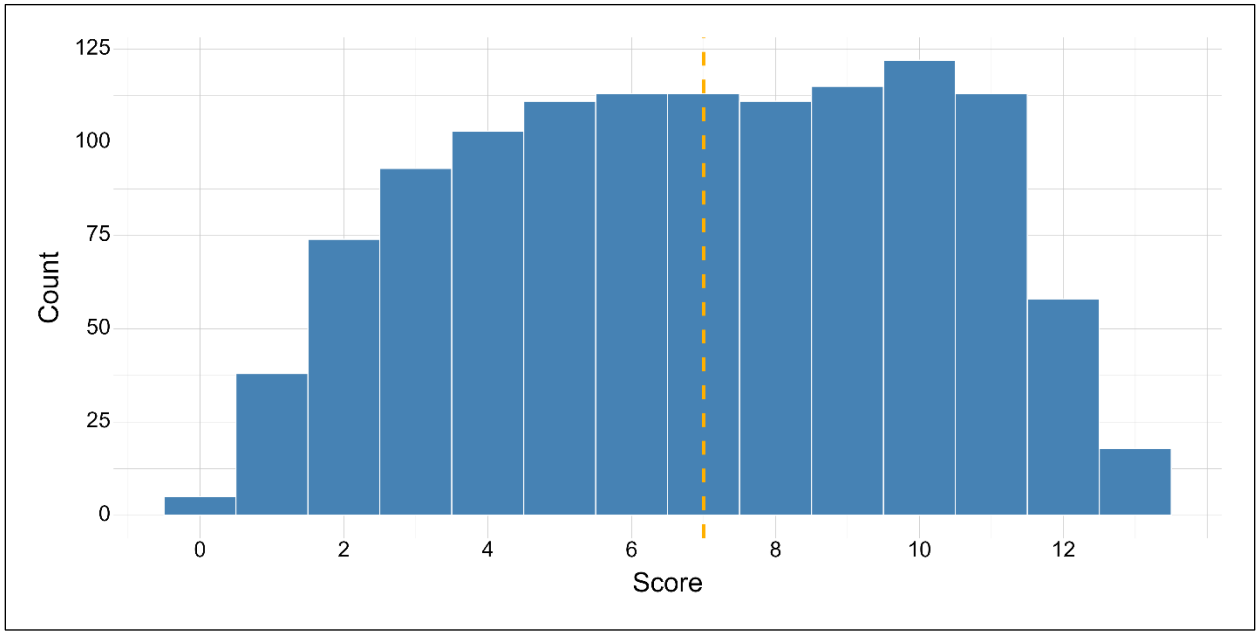


Figure 36.

*Distribution of Scores in Geometry and Measurement Domain (Total Possible = 8 Points)*

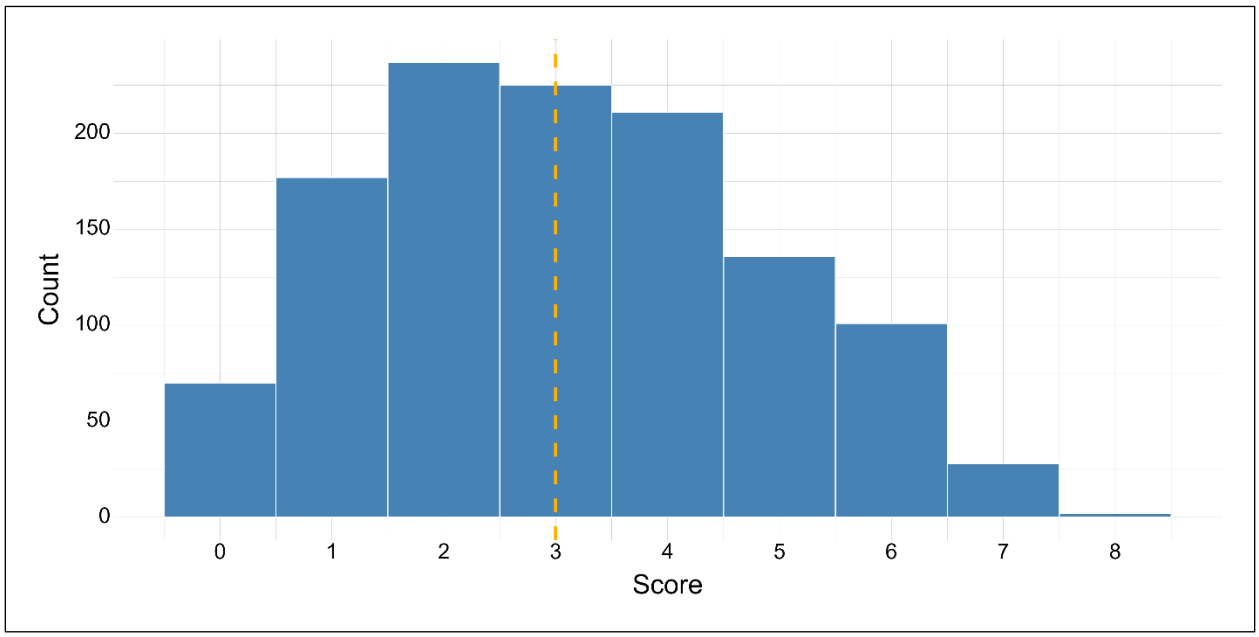
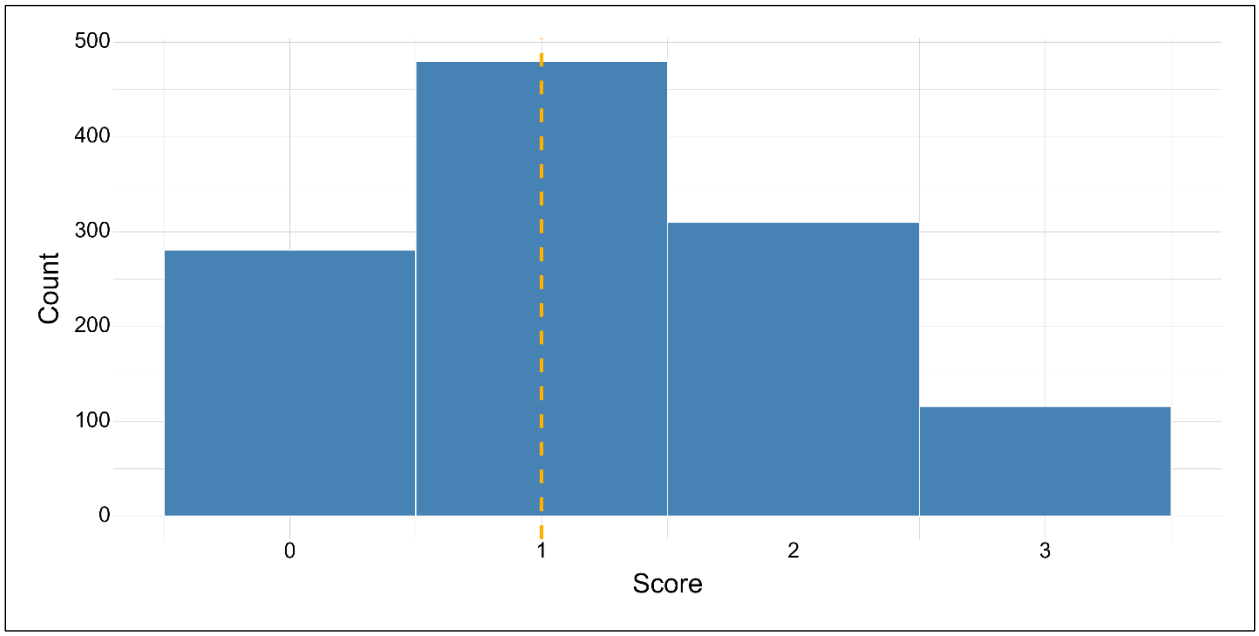



Figure 37.

*Distribution of Scores in Data and Statistics Domain (Total Possible = 3 Points)*




# Validity Argument

One of the goals of the PIE project was to design an approach to evaluating technical adequacy, including scoring model, theory of action, and validation plan that supports multiple uses of the

system. We started by defining a theory of action (see Figure 1) and supporting propositions to serve as the basis of an argument-based approach to validation. We used the theory of action to define the breadth of validity studies necessary to evaluate an integrated assessment model for summative assessment for use in state accountability models under the Every Student Succeeds Act (2015), and define and evaluate a scoring model to produce instructionally useful and summative achievement results. The PIE theory of action and validation plan is used to outline the evidence needed to support valid interpretations of assessment results used for both formative and summative purposes.

The scope of evidence provided in the previous chapters of this report are most relevant to claims D, E, F, G, I, and J. This chapter summarizes evidence from additional resources, including the technical report *PIE as a Proof of Concept for Future Summative Use: Evidence From a Pilot Study* (ATLAS, 2025d), which describes summative scoring models and additional methods and approaches for future statewide summative uses of a PIE system.

For each statement in the theory of action, we summarize the underlying propositions and associated evidence informing evaluation of the proposition, indicate the type of evidence (e.g., content, response process, consequences), and identify the source (either a chapter within this report or separate source) containing a full description of the evidence.

## Design

Design components of the PIE assessments include learning pathways, content standards, items/tasks, teacher resources, and training. These propositions each serve as inputs into delivery claims for PIE assessments.

*Claim A: Learning pathways accurately describe typical development pathways of knowledge, skills, and understandings.*

| Proposition | Evidence | Source | Type |
|---|---|---|---|
| Learning pathways (cognitive learning model) accurately describe the development of knowledge and skills. | Description of learning pathway development and literature basis, expert review | Kim, E. M., et al. (2024). | Content |
| Pathway levels reflect the critical knowledge, skills, and understandings that precede and include those required by the grade-level content standards. | Description of learning pathway map development process including literature basis, expert review, and DESE review; empirical data | Kim, E. M., et al. (2024).<br>Chapter 5 of this report | Content |

For each of the 25 fifth-grade mathematics content standards included in the PIE project, a learning pathway was constructed to portray a typical learning path toward KSUs that the standard targeted and to locate groups of the most critical KSUs that could serve as assessment targets. The full development process comprised (a) standard analysis, (b) literature synthesis, (c) progression cross-checking, (d) pathway drafting, and (e) expert review. The development process resulted in 213 KSUs, 476 KSU relationships, and 75 pathway levels. The process of synthesizing learning literature, as well as other related resources, produced a solid literature base for the learning-pathway construction.

The pathway level expert review, performed by two panels of four educators each, provides evidence for the reviewed pathway levels. The panels believed the large percentages of levels satisfied the five criteria (93.3%, 68.0%, 87.5%, 70.7%, and 57.3% for the consistency, progression, size, alignment, and clarity criteria, respectively). The final learning pathways were approved by MO-DESE.

This technical report describes the student-mastery ranges of crucial KSUs by tested pathway level (20.6%–79.6% for Level 3, 48.1%–97.6% for Level 2, and 66.1%–99.4% for Level 1). These reported data on the level mastery support the tested pathway levels' hierarchical relationship.

*Claim B: Assessment tasks are designed to measure the knowledge and skills delineated in the learning pathways, including higher-order skills.*

| Proposition | Evidence | Source | Type |
|---|---|---|---|
| Pathway levels are specified at an appropriate grain size for assessment. | Description of learning pathway level development, model fit | Kim, E. M., et al. (2024).<br>Chapter 5 of this report<br>ATLAS. (2025a). | Content, Internal structure |
| Assessment tasks are aligned to a level in the learning pathway. | Task models, item-writing handbook, external review checks for alignment, item analysis, modeling results | ATLAS. (2025a).<br>Chapter 5 of this report | Content, Internal structure |
| Assessment tasks elicit the cognitive skills specified in the intended learning pathway level. | Cognitive labs, task models, item-writing training and guidelines, internal and external review | ATLAS. (2025a).<br>ATLAS. (2025c). | Response process, Content |
| Assessment tasks are free of extraneous content. | Task models, item-writing training and guidelines, internal & external review, field test and pilot study data | ATLAS. (2025a).<br>Chapter 1 of this report | Content |
| Assessment tasks do not contain content that is biased against or insensitive to subgroups of the population. | Item-writing training and guidelines, internal and external review, DIF analyses and content review | ATLAS. (2025a).<br>Chapter 1 of this report | Content, Internal structure |
| Assessment tasks elicit consistent response patterns across different administration formats (e.g., different accessibility features turned on/off). | Task models, item-writing training and guidelines, accessibility manual, internal and external review, *p*-value, standardized difference values<br>Comparability studies (*future study*) | ATLAS. (2025a).<br>Chapter 1 of this report | Response process |

Development of the LPs involved constructing a progression of relevant KSUs that lead to the learning target and selecting essential KSUs for the three pathway levels for each standard. As documented in Kim et al. (2024), the five-step process for learning-pathway development aimed to ensure the three vertical pathway levels increased in cognitive complexity as the levels progressed and included distinct KSUs that are logically sequenced and appropriately sized for assessment purposes. Eight educators participated in an expert review of the LPs. Expert reviewers used specific content and structure criteria to evaluate whether the content of each LP included KSUs that were appropriately sized and different from neighboring pathway levels. More than 85% of the LPs met the review criteria for appropriate grain size. Feedback from the external review was used to inform revisions to the LPs to ensure appropriate grain size for all pathway levels. In addition, an empirical validation of the LPs, which was conducted using data from the 2024 PIE field test, concluded that pathway levels were specified at an appropriate grain size for assessment; 80% of the calibrated models, fit individually for each content standard, demonstrated adequate absolute fit. For additional information, refer to the *PIE Assessment Design and Development Technical Report* (ATLAS, 2025a).

Each assessment task measures the KSUs associated with one learning pathway level. Assessment tasks were designed to elicit evidence of the targeted cognition associated with the learning pathway level. Prior to item development, the math content-development team created task models that described the targeted cognition for each of the 75 pathway levels. The *PIE Assessment Design and Development Technical Report* (ATLAS, 2025a) describes the processes for task-model and item development. Each task model included information about the targeted cognition, including types of evidence related to the KSUs for the LP Level statement, as well as KSUs that are considered outside the targeted cognition. Task models also included Universal Design for Learning and accessibility considerations, which provided guidelines for constructing assessment items that are accessible to a range of test takers and do not present any barriers for students to demonstrate KSUs associated with the level statement. Evidence from item analyses confirms that assessment tasks are aligned to a level in the learning pathway. *P*-values demonstrated that students classified as masters were more likely to succeed on aligned items than were nonmasters. DIF analyses indicated that items functioned consistently across student groups matched on ability, providing evidence of fairness and alignment. Last, comparing item twins with their content-equivalent counterparts showed consistent performance patterns.

Information from the task models was used throughout the item-development process—including item writing, internal review, and external review—to ensure that the KSUs focused on the targeted cognition of the level statement and assessment tasks elicited evidence of the targeted cognition. When developing items, different administration formats were considered, such as TTS, so students could engage with the content in a way that allowed them to demonstrate their understanding of the skills. Internal and external reviewers provided feedback about the assessment tasks using specific content and fairness criteria. The content criteria focused on whether the items aligned to the level statement and provided evidence of the KSUs defined by the level statement. The

fairness criteria addressed any concerns with the items related to bias, sensitivity, and accessibility. All content and fairness comments were resolved, and additional reviews of the items were conducted to verify no content or fairness concerns remained.

Student responses to PIE assessment items should reflect their KSUs rather than construct-irrelevant factors. PIE items were designed to minimize construct-irrelevant variance and elicit the intended cognition required for the task. Cognitive labs offer some evidence that PIE items do not present unintended barriers to the intended response process caused by construct-irrelevant features or item-response demands. Of the nine PIE item types administered during the cognitive labs, the majority were observed to elicit either the intended response process or a student misconception unrelated to the features of the item. Overall, students were able to interact with the prototype items, which included matching, constructed response, and interactive features such as clickable hot spots, as intended; one item type (i.e., drop-down) generated more unintended response processes than others did. However, because the cognitive-lab test forms included only one drop-down item, any interpretations of this finding should be made with caution.

Assessment tasks should be free of biased or sensitive content to ensure that results reflect student knowledge and skills rather than irrelevant factors. The PIE development process incorporated safeguards at multiple stages. The item-writing handbook provided explicit guidance to item writers to avoid content that could disadvantage subgroups. After development, items underwent internal and external reviews focused on fairness, sensitivity, and accessibility, as well as presentation checks in the Student Portal. Statistical analyses provided additional evidence. DIF analyses were conducted for gender and race subgroups to evaluate whether items functioned differently for students of similar ability. About 10%–12% of items were flagged, but only a small number (two to four items per analysis) showed moderate or large effect sizes. Together, the design safeguards, review processes, and DIF results provide evidence that PIE assessment tasks were developed and evaluated to minimize bias-and-sensitivity concerns, supporting the proposition that tasks do not contain content that is unfair to subgroups of the population.

Assessment tasks should elicit consistent response patterns across different administration formats. Several design features support this expectation. Task models were written at a grain size that ensured items included all of the necessary evidence required in each pathway level statement, regardless of item type. The item-writing handbook and accessibility manual provided procedures to guide item writers in developing tasks aligned to the intended pathway levels while maintaining consistency across formats. Empirical evidence from the 2024–2025 pilot study further supports this claim. Across 476 administered items, roughly 86%–97% of items fell within the acceptable ranges of item difficulty for both masters and nonmasters. In addition, 145 item twins, constructed to be content-equivalent to original items but with varied contexts or values, were evaluated. Over 90% of effect sizes for both masters and nonmasters were small or moderate, indicating that item twins functioned similarly to their counterparts. Overall, the combined design specifications, review processes, and pilot study results provide evidence that PIE assessment tasks function consistently across administration formats, ensuring that student scores are not influenced by the system features in use.

*Claim C: PIE assessments and the system used to deliver PIE assessments, are designed to maximize student accessibility.*

| Proposition | Evidence | Source | Type |
|---|---|---|---|
| The learning pathway levels that assessments measure provide access to content for the full range of students. | Description of learning pathway development and literature basis, expert review; empirical data and teacher feedback from pilot study | Kim, E. M., et al. (2024).<br>Chapter 5 of this report<br>ATLAS. (2025b). | Content |
| Assessment tasks are designed to be accessible to students. | Task models, cognitive labs, item-writing training and guidelines including UDL guidelines, item-writer qualifications, internal and external review, teacher survey | ATLAS. (2025a).<br>ATLAS. (2025c).<br>Chapters 4 and 5 of this report | Content |
| System design is consistent with accessibility guidelines, APIP and WCAG standards, and contemporary code. | System documentation and evaluation | ATLAS & DESE. (2024a).<br>Chapter 4 of this report | Content |
| Supports needed by the student are available within and outside of the assessment system. | System documentation, teacher survey | ATLAS & DESE. (2024a).<br>Chapter 4 of this report | Content |
| Item types support the range of students in presentation and response. | Item-writing training and guidelines, cognitive labs, internal and external review | ATLAS. (2025a).<br>ATLAS. (2025c). | Response process, Content |

The development process and procedures for learning pathways (including cross-checks with existing learning map models that serve a broad range of students) supported creation of LP levels that reflect how students learn the KSUs leading up to grade-level academic expectations. Within the scope of the project, the LPs were developed to measure two levels of precursor KSUs to support students who were struggling to meet grade-level academic expectations. Findings from the pilot study indicated that the majority of students (66.1%–99.4%) were able to demonstrate mastery of Level 1 skills, nearly half or more students (48.1%–97.6%) demonstrated mastery of Level 2 skills, and approximately 20.6%–79.6% of students demonstrated mastery of Level 3 skills. These

results suggest that the precursor LP levels provided access to the content for more students than did the grade-level content standard alone (i.e., Level 3). However, some teachers noted that, for some of their higher achieving students, the LPs and assessments were less useful given their focus on precursor KSUs. Future development efforts could focus on creating a fourth LP level that extends beyond the grade-level standard.

Assessment items are designed to be accessible to students in multiple ways. First, the items must appropriately measure their intended LP level. The evidence-centered, design-based PIE task models for each LP level support item writing by specifying the item-level metadata, targeted cognition, recommended item types, Universal Design considerations, and test-administration accessibility considerations. Item-writing training and guidelines provided guidance to item writers for how to write items measuring the LP level. After the item-writing workshop with educators, the draft items were further developed by the PIE project team; development steps included: ensuring alignment to the level statement; conducting accessibility reviews to identify any accessibility, bias, or sensitivity concerns; and editorial review. Additional reviews by MO-DESE project staff and educators also checked for alignment and accessibility considerations. Finally, most teachers reported that they believed the content of PIE assessment was accessible to students (78%), and student responses to PIE items accurately reflected their KSUs (69%).

For the Kite Suite to maximize accessibility, system design should be consistent with accessibility guidelines, accessible portable item protocol (APIP) standards, web content accessibility guidelines (WCAG), and current best practices. For the Kite Suite to be fully accessible to all students, all supports needed by students within and outside of the assessment system should be available. Test administration documentation describes the types and range of accessibility supports available. Furthermore, the majority of teachers (87%) agreed that students had access to all accessible supports necessary to participate in the assessment,

System accessibility also means that item types support the full range of students in presentation and response. Multiple item types allow students to demonstrate their KSUs. Cognitive labs showed that students can respond to a range of item types, including constructed-response and technology-enhanced items such as gap-match, hot-spot, matching-lines, and matrix-interaction items; however, some students struggled more with drag-and-drop items. Overall, students were able to interact with the prototype items as intended; one item type (i.e., drop-down items) generated more unintended response processes than others did. However, because the cognitive-lab test forms included only one drop-down item, any interpretations of this finding should be made with caution. These findings were used to inform the number and types of interactions students need to engage in for drag-and-drop items included in the pilot study.

Finally, the Kite Suite must also be accessible to educators. Test administration documentation describes how educators use Educator Portal, including the Content Groups page, to create customized content groups according to their assessment and instruction plan, initiate test assignments for their students within their instructional cycles, access mastery results, and manage student data. The

documentation also describes how test administrators use Student Portal to administer assessments to students. Teacher focus groups informed the design of the Content Groups page and reporting dashboard. Most pilot study teachers (85%) thought the PIE content-groups-management page in Kite Educator Portal was easy to navigate, and 87% agreed that the reporting dashboard was easy to navigate.

*Claim D: Training and resources support educators in using the PIE system to inform instruction.*

| Proposition | Evidence | Source | Type |
|---|---|---|---|
| Training is designed to build educator knowledge and skills for creating groups of content standards for assessment. | Summary of training, training-module focus groups, module self-assessment data, teacher survey | Chapter 4 of this report<br>ATLAS. (2025b).<br>Nash et al. (2025). | Consequences |
| Training is designed to strengthen educator knowledge and skills for planning and implementing cycles of instruction and assessment. | Summary of training, training-module focus groups, module self-assessment data, teacher survey | | Consequences |
| Resources support educators' understanding of learning pathway content. | Training-module focus groups, teacher survey, | | Consequences |
| Training and resources are designed to support educators' data-based instructional practices. | Summary of training, training-module focus groups, teacher survey | | Consequences |

PIE training was designed to prepare educators for working with the PIE system components, including the use of LPs to develop customized PIE assessments, creating content groupings using priority standards, providing instruction on content groups, administering the PIE assessments throughout the instruction and assessment cycle, and using PIE assessment results and data to inform instruction.

Training should strengthen educator knowledge and skills for using the PIE system and assessing students. System documentation, including user manuals, a summary of the PIE training and descriptions of the content of the e-learning modules, describes the approach used to train teachers on how to use the PIE system.

Teachers were required to complete training to participate in the pilot study. With the exception of two modules (i.e., *Interpreting PIE Assessment Results* [69%] and *Using PIE to Inform Your Instruction* [80%]), the participation rate for the other modules was at or near 100%.

Data collected from educators offered insight into how teachers used PIE training to plan instruction and administer assessments. Over 95% of teachers agreed or strongly agreed that, overall, the PIE training and resources prepared them to create content groups for instruction and assessment. Approximately 85% of educators agreed or strongly agreed that PIE training modules and associated resources provided the requisite knowledge and skills to support implementing cycles of instruction and assessment. Over 75% of educators agreed or strongly agreed that PIE training and resources prepared them to interpret and use PIE reports and data to inform their instruction.

## Delivery

Delivery claims pertain to the combined set of administered testlets, student interactions with the system, and educator instruction and administration. These propositions each serve as inputs into scoring claims for PIE assessments.

*Claim E. Educators provide instruction aligned with the learning pathways to students.*

| Proposition | Evidence | Source | Type |
|---|---|---|---|
| Educators provide students the opportunity to learn content aligned with the learning pathways—including grade-level content standards—using strategies appropriate for a continuum of learners. | Test-administration procedures, teacher survey, educator selection & plan creation patterns | ATLAS & DESE. (2024b).<br>Chapters 2 and 4 of this report<br>ATLAS. (2025b). | Content |
| Educators have access to, and receive training on, high-quality instructional materials and/or practices. | To be evaluated as a future study | | |

Educators should provide students with the opportunity to learn the content in the standards-aligned LPs. The *PIE Pilot Study Educator Portal User Guide* (ATLAS & DESE, 2024b) and *Test Administration Manual* (ATLAS & DESE, 2024a) describe expectations that educators provide aligned instruction and determine when students are ready for assessment based on their opportunity to learn the relevant content.

Survey data indicated that 83% of teachers agreed that students had the opportunity to learn content aligned with the PIE LPs, and 77% reported being able to use instructional strategies to reach a continuum of learners at various places in the learning pathways.

In focus groups, teachers generally agreed that the PIE system was a valuable resource to gauge student understanding, identify content to reteach, inform instructional small groups, and plan future lessons. However, several participants expressed concerns about the length of the assessment window. Some educators, however, felt rushed and unable to instruct and assess all the content within the window because of the amount of content in the PIE Learning Pathways. As a result, only 26 (1.7%) students participating in the pilot study completed all 25 standards outlined in the assessment blueprint during the instructionally embedded assessment window.

These results collectively indicate variability in the amount of instruction for the full breadth of academic content, with opportunity for continuous improvement. Extending the assessment window and continuing to support educators in understanding the alignment between local curriculum standards and state academic content standards may help address some of these challenges.

*Claim F. Educators administer assessments consistent with intended design.*

| Proposition | Evidence | Sources | Type |
|---|---|---|---|
| Educators plan and implement cycles of instruction and assessment consistent with intended design. | Teacher survey, teacher use of resources, educator plan creation patterns | Chapters 4 and 5 of this report | Content |

Educators should administer the instructionally embedded assessments as intended. According to the *PIE Pilot Study Educator Portal User Guide*, *Test Administration Manual*, and *PIE Pilot Study: Design and Administration Evidence*, intended administration includes

1. selecting standards for content groups (i.e., planning)
2. administering the baseline assessment
3. evaluating baseline results
4. teaching
5. administering the midway assessment
6. evaluating midway results
7. teaching
8. administering the end-of-unit assessment
9. evaluating end-of-unit results
10. providing additional instruction as needed

The intended administration was also described in the required training modules that teachers completed at the start of the pilot study.

Survey data indicated that 98% of pilot study teachers agreed that the PIE training modules and resources/job aids helped them know how to create content groups for assessment and instruction, and 83% believed it provided them with the knowledge and skills to implement cycles of assessment and instruction. Similarly, 79% of teachers indicated they were able to administer assessments at instructionally relevant points in time (i.e., after instruction on the assessed content had occurred), and 74% indicated they could use the PIE assessment results to plan next steps for their instruction. However, a smaller percentage of teachers (62%) thought the PIE assessment tasks were aligned with instruction, and only 56% agreed that the length of the assessment window was sufficient for completing the instruction and assessment cycle for each of their planned content groups. Furthermore, given the varied blueprint

coverage and assessment-completion rates, there are opportunities to improve the system to support intended implementation practices.

*Claim G. Students interact with the PIE assessments to demonstrate their knowledge and skills.*

| Proposition | Evidence | Source | Type |
|---|---|---|---|
| Students are able to respond to tasks irrespective of a sensory, mobility, health, communication, or behavioral constraint. | Test-administration procedures, teacher survey, item and test completion rates<br>Aberrant-response analyses (future study) | ATLAS & DESE. (2024a).<br>Chapter 4 of this report | Response process |
| Student responses to items reflect their knowledge, skills, and understandings. | Teacher surveys, cognitive labs<br>Aberrant-response analyses (future study) | ATLAS & DESE. (2024a).<br>ATLAS. (2025c).<br>Chapter 4 of this report | Response process |
| Students are able to interact with the assessment tasks as intended. | Test-administration procedures, cognitive labs, teacher surveys | | Response process |

To demonstrate their KSUs within the PIE system, students must have the opportunity to respond to tasks without limitations related to sensory, mobility, health, communication, or behavioral needs. Test-administration guidelines specify the available accessibility supports and outline procedures to ensure assessments are delivered in ways that meet individual student needs. Evidence from the teacher survey supports this claim: 87% of teachers agreed that students had access to the supports necessary to participate, and 78% agreed that the content of the PIE assessment was accessible to their students. Test-administration data showed that patterns of item and test completion rates demonstrated broad participation across standards, although only 1.7% of students completed all 25 standards. However, these outcomes largely reflected teachers' administration practices rather than students' access barriers.

Student responses to PIE assessment items should reflect their KSUs rather than construct-irrelevant factors. PIE items were designed to minimize construct-irrelevant variance. Cognitive labs offer some evidence that PIE items support the intended response process; that is, they do not present unintended barriers caused by construct-irrelevant features or item-response demands. Of the nine PIE item types administered during the cognitive labs, the majority were observed to elicit the intended response process or elicited a student misconception unrelated to the features of the item. For more information on PIE cognitive labs, refer to the technical report, *Evaluating Pathways for Instructionally Embedded Assessment (PIE) Assessments: Evidence From Cognitive Labs*

(ATLAS, 2025c). Additionally, on the 2025 teacher survey, 69% of educators agreed that students' responses to PIE items accurately reflected their KSUs.

Students are expected to interact with the PIE system in ways that are consistent with its intended design, as outlined in administration documentation specifying permitted and prohibited practices. Evidence from the PIE cognitive labs indicates that students were able to interact as intended with the majority of prototype items, which included matching, constructed response, and interactive features such as clickable hot spots; one item type, drop-down items, generated more unintended response processes than others did. However, because the cognitive-lab test forms included only one drop-down item, any interpretations of this finding should be made with caution. The teacher-survey results also reinforced the claim that students were able to access and interact with the assessment tasks as intended. Teachers said students generally responded positively to the assessment format, noting benefits such as the untimed administration, and alignment with classroom instruction. At the same time, teachers identified areas for improvement, including a need for additional tools (e.g., read-aloud options) and interface refinements.

*Claim H. Assessments administered multiple times at instructionally relevant points in time and at the end of year, measure students' knowledge, skills, and understandings.*

| Proposition | Evidence | Sources | Type |
|---|---|---|---|
| Educators administer instructionally embedded assessments that cover the blueprint standards. | Blueprint coverage and completion rates during instructionally embedded window, educators' selection patterns of standards for content groupings | Chapter 4 of this report | Content |
| Educators administer the end-of-year assessment that covers the blueprint standards. | End-of-year assessment-completion rates | Chapter 4 of this report | Content |
| The collection of administered assessments measure the breadth and depth of the KSUs in the Missouri standards.[1] | Test blueprint, blueprint coverage and completion rates, description of learning pathway development, expert review | Chapter 4 of this report<br>Kim, E. M., et al. (2024). | Content |

For administered PIE assessments to capture the intended breadth of skills outlined in the learning map, coverage of the full blueprint is essential. The PIE assessment blueprint for the 2024–2025 instructionally embedded window included 25 content standards across four fifth-grade domains. Procedures recommended educators to create content groups of two to five standards for assessment, although content groups observed in practice contained between one and seven standards. The PIE assessment blueprint for the end-of-year assessment window simultaneously covered all 25 standards, with 50 items (two per standard) administered via two test forms.

Fifty-five educators across 69 classrooms participated in the instructionally embedded window. In total, 323 content groups were created, the majority (78.9%) of which aligned with the recommended range of two to five standards per group. A smaller share of groups (21.2%) included only one standard or six or more standards, representing departures from the recommendations.

[1] For the purposes of the PIE grant-funded project, 25 Missouri fifth grade mathematics standards were prioritized for development. Future development would include the full set of learning standards.

Classroom-coverage data showed that most classrooms engaged with a substantial portion of the blueprint, though full coverage was uncommon. About one-third of classrooms completed assessments for 11–15 standards, while smaller proportions completed fewer than 10 or more than 20 standards. Only one classroom (1.4%) completed all 25 standards during the window. These findings indicate that educators generally administered instructionally embedded assessments in ways that provided broad representation of the blueprint standards, although complete coverage across all 25 standards was rare.

Forty-six educators across 55 classrooms participated in the end-of-year assessment window and administered the end-of-year assessment aligned to the 25 fifth-grade content standards. Completion data indicated that a vast majority of classrooms administered the final assessment. Together, the instructionally embedded and final assessments measured the breadth and depth of the KSUs in the Missouri Standards.

## Scoring

Scoring claims included mastery determinations, overall achievement, and usability of results. These propositions each served as inputs into the intended outcomes of PIE assessments.

*Claim I. Mastery results represent what students know and can do relative to the learning pathways.*

| Proposition | Evidence | Source | Type |
|---|---|---|---|
| Learning pathway level mastery classifications are reliable. | Reliability analyses | Chapter 5 of this report | Internal structure |
| Mastery status reflects students' knowledge, skills, and understandings. | Scoring method, model fit, model parameters | Chapter 5 of this report | Internal structure |
| Mastery results are consistent with other measures of students' knowledge, skills, and understandings. | Teacher survey | Chapter 5 of this report | Other measures |

At the pathway level, 49 (65.3%) pathway levels met the recommended 0.83 cutoff for fair classification accuracy, and 66 (83%) pathway levels met the recommended 0.70 cutoff for fair classification consistency. These findings indicated that a vast majority of

pathway level classifications were reliable and exhibited low measurement error. The evidence was consistent across pathway levels, indicating that measurement precision was moderately stable across the continuum of KSUs represented by the three levels.

At the model level, 80% of the estimated learning pathway level models showed acceptable levels of absolute model fit. At the item level, the quality of administered items was evaluated, and items demonstrating poor fit were flagged according to defined criteria. In total, 90.9% of the calibrated items exhibited an average score for masters that was greater than 0.5 times the maximum number of possible points, and 78.2% of items exhibited an average score for nonmasters that was less than 0.5 times the maximum number of possible points. Together, the evidence largely indicated that mastery status reflected students' KSUs and highlights opportunities for additional collection of data.

Last, evidence from the teacher survey indicated that most teachers (72%) agreed or strongly agreed that mastery results were consistent with their own perceptions of their students' mastery of assessed KSUs. This agreement between teacher judgment and assessment outcomes supported the validity of mastery results, while also highlighting the need for additional sources of evidence to further strengthen this claim.

*Claim J. Instructionally embedded results provide instructionally useful information.*

| Proposition | Evidence | Source | Type |
|---|---|---|---|
| Reports are fine-grained. | Development process, teacher focus groups, reporting dashboard | Chapter 5 of this report<br>Nash et al. (2024). | Consequences |
| Instructionally embedded score reports are instructionally useful, and provide relevant information for educators. | Pilot study focus groups and interviews, teacher survey | ATLAS. (2025b).<br>Chapter 5 of this report | Consequences |

Initial prototypes of the reporting dashboard were developed and iteratively refined using feedback from the Project Advisory Committee (PAC) and focus groups comprising 11 Missouri educators, including feedback about grain size and potential uses of the results. Feedback informed the final prototypes, which were used to build the online PIE reporting dashboard.

Focus group and interview session participants described using PIE results and the PIE reporting dashboard to make instructional adjustments and plan for next steps in their instruction. On the teacher survey, most teachers agreed or strongly agreed that the PIE score reports provided instructionally relevant information (74%) and could be used to plan for next steps in their instruction (79%).

*Claim K. Summative results accurately reflect student achievement of grade-level academic content standards.*

| Proposition | Evidence | Source | Type |
|---|---|---|---|
| Summative results reflect multiple opportunities for students to demonstrate what they know and can do. | Description of scoring model and model-evaluation data | ATLAS. (2025d). | Internal structure |
| Summative results are correlated with other measures of student achievement. | Correlational analyses (future study) | N/A | Internal structure |
| Summative scores are reliable. | Reliability analyses | ATLAS. (2025d). | Internal structure |

Several item response theory (IRT), DCM, and hybrid IRT/DCM scoring models were evaluated; these scoring models derived summative scores from response information provided by students across multiple testing opportunities within the instructionally embedded window and, optionally, the end-of-year assessment. Results from the model-fit analyses indicated that versions of a summative model from all three frameworks (i.e., IRT, DCM, hybrid) demonstrated adequate model fit. Thus, the selection of a future summative model should be driven by consistency with other claims in the theory of action, intended uses of the system, and intersections with other components of an overall assessment system (e.g., standard-setting methods, growth models).

For each summative model that was estimated, we also evaluated the reliability of the resulting scores. For the IRT and hybrid models, reliability was evaluated using the conditional standard error of measurement. For the DCM models, reliability was defined as the classification consistency. All models demonstrated high levels of reliability. Only DCM mastery classifications from the end-of-year assessment showed low levels of reliability, but this finding was primarily driven by the design of the end-of-year assessment (i.e., two items per standard), rather than properties of the content or estimated model.

## Outcomes

The intended outcomes of PIE assessments are summarized.

*Claim L (Proximal). Educators make instructional decisions based on data from the PIE assessments.*

| Proposition | Evidence | Source | Type |
|---|---|---|---|
| Educators are trained to interpret assessment results as intended. | Training module focus groups and self-assessment data, teacher survey, pilot study focus groups and interviews | Chapter 4 of this report<br>ATLAS. (2025b). | Consequences |
| Educators interpret assessment results as intended. | Training module focus groups and self-assessment data, reporting dashboard focus groups, teacher survey, pilot study focus groups and interviews | | Consequences |
| Educators are trained to use assessment results to inform next steps in instruction and instructional activities. | Training module focus groups and self-assessment data, reporting dashboard focus groups, teacher survey, pilot study focus groups and interviews | | Consequences |
| Educators use assessment results to inform next steps in instruction and instructional activities. | Reporting dashboard focus groups, teacher survey, pilot study focus groups and interviews | | Consequences |
| Educators use assessment results as one source of evidence of student learning. | Teacher survey, pilot study focus groups and interviews | Chapter 5 of this report<br>ATLAS. (2025b). | Consequences |
| Educators reflect on their instructional decisions. | Pilot study focus groups | ATLAS. (2025b). | Consequences |

To make informed instructional decisions, educators should be trained to interpret results as intended and be able to use PIE results to inform their instruction. Documentation describing the design and implementation of PIE e-learning modules provided evidence that educators received training on using results to inform their instruction. Focus groups conducted with educators after completion of the training modules and module self-assessment data offered evidence that educators were trained to interpret PIE assessment results as intended. Interview sessions and focus groups conducted at the conclusion of the PIE pilot study provided additional sources of evidence to support the claim that teachers received adequate training to interpret PIE assessment results appropriately. Additional PIE training, including a module called “Driving Your Instructional Decision-Making With Data”, provided educators with

guidance on how to use results to inform next steps in instruction. On the teacher survey, over 75% of educators agreed or strongly agreed that the PIE e-learning modules and resources helped them understand how to interpret and use data to inform their instruction.

For educators to make sound, data-based instructional decisions, they must interpret assessment results as intended and be able to use those results as one source of information about student learning to guide future instructional decisions. PIE system documentation, including information about the PIE LPs and the *PIE Pilot Test Administration Manual* (ATLAS & DESE, 2024a), describes intended uses of PIE results. Evidence collected from participants in the pilot study, including feedback from teachers who took part in interviews and focus group sessions, described how educators used PIE results to make instructional decisions and changes in their instructional approach. Educators described ways that they used PIE results to pinpoint areas in which students would benefit from targeted instruction, and the specific adjustments they made, such as developing visual models or incorporating physical manipulatives into lessons. On the teacher survey, most teachers (89%) agreed or strongly agreed that the PIE score reports provided one source of evidence of student learning.

Findings from the PIE focus groups and interviews indicated that, while educators were able to use their training to interpret results and then use those results to make data-based instructional decisions, they did not discuss ways in which they reflected on their instructional decisions. This finding represents an opportunity for future research into ways to support educators in making regular reflection a part of their routine instructional practice.

*Claim M (Proximal). Students make progress toward mastery of grade-level content standards.*

| Proposition | Evidence | Source | Type |
|---|---|---|---|
| Students demonstrate academic progress toward grade-level content standards. | Pilot study assessment data | Chapter 5 of this report | Consequences |

Students should demonstrate academic progress toward grade-level expectations. The design of the PIE system supported this intention through the map structure of LPs, which embedded rigorous, grade-level standards and guided aligned instruction. Students demonstrated progress by engaging with assessment tasks to show their KSUs and by accessing opportunities to retest on skills not previously mastered.

Mastery rates reflected the expected differentiation across pathway levels: Level 1 skills showed higher mastery rates (66.1%–99.4%), Level 2 skills were more variable (48.1%–97.6%), and Level 3 skills, aligned to grade-level expectations, were the most challenging

(20.6%–79.6%). These outcomes indicated that the system captured meaningful differences in student performance across levels of complexity in content. Retest opportunities provided additional evidence of progress. Among students who did not initially demonstrate mastery, 19.0%–74.5% achieved mastery upon retest of Level 1 skills, and 7.8%–77.3% did so for Level 2 skills. Overall, the observed mastery rates and retest outcomes supported the proposition that students demonstrated academic progress toward grade-level content standards within the PIE assessment system.

*Claim N (Distal). State and district education agencies use results for monitoring and resource allocation.*

| Proposition | Evidence | Type |
|---|---|---|
| District and state staff use aggregated information to evaluate programs and adjust resources. | To be evaluated as a future study | Consequences |

*Claim O (Distal). Parents understand and use results to monitor their student's learning.*

| Proposition | Evidence | Type |
|---|---|---|
| Parents have access to their student's results from instructionally embedded and end-of-year assessments. | To be evaluated as a future study | Consequences |
| Parents understand how to interpret assessment results provided in reports. | To be evaluated as a future study | Consequences |
| Parents use assessment results to monitor their student's learning throughout and at the end of the year. | To be evaluated as a future study | Consequences |

*Claim P (Distal). Student achievement improves.*

| Proposition | Evidence | Type |
|---|---|---|
| Over time, the percentage of students who meet proficiency standards by | To be evaluated as a future study (using longitudinal data) | Consequences |

| | | |
|---|---|---|
| the end of the school year will increase. | | |

# Conclusions

Two goals of the PIE project were addressed by the pilot study and the data that were collected as part of the study. Goal 2 was to evaluate the usability of the PIE assessment system under natural conditions. Overall, the intended outcomes of Goal 2 were met. Table 28 summarizes the objectives and outcomes relevant to Goal 2.

Table 28.

*Objectives and Project Outcomes for Goal 2*

| Objective | Objective outcomes | Project outcomes |
|---|---|---|
| **Objective 2.1**<br>Build an instructionally embedded, general-education assessment system to support online delivery of PIE. | Design and develop key components of the PIE system, including a teacher user-interface tool and reporting dashboard. | Initial prototypes for the teacher user-interface tool for creating and assigning assessments and the reporting dashboard were developed and iteratively refined according to feedback from members of the Project Advisory Committee (PAC) and focus groups with 11 experienced Missouri educators. The final prototypes were used to build the online PIE platform.<br><br>The Kite technology platform used to deliver and score the assessments was further developed to support automated scoring using DCM. Specifically, hierarchical DCMs were calibrated for each LP, and model parameters were used in the system to automatically score and report students' mastery statuses for each LP level. These mastery results were populated in the reporting dashboard (via classroom mastery report and individual student report). |

| Objective | Objective outcomes | Project outcomes |
|---|---|---|
| **Objective 2.2**<br>Design and develop training and assessment literacy resources for educators on purposes and uses of the PIE assessments. | Design and develop a professional-learning solution that prepares teachers to effectively use the PIE system by building their proficiency in creating assessments, administering them, and interpreting results to inform instruction. | A set of six self-paced e-learning modules and corresponding job aids were developed and refined using feedback from focus groups composed of nine Missouri teachers and the PAC.<br><br>The modules introduced PIE system components, developed teacher assessment literacy, provided a walk-through of the administration of PIE, oriented teachers to PIE assessment reports, and provided an opportunity for teachers to interpret data and explore implications for classroom teaching via a model of data-based instructional decision-making. Teachers were provided with a set of job-aid resources that reinforced module content and could be used as quick-access reference materials. |
| **Objective 2.3**<br>Conduct a pilot study of the PIE system. | Recruit and prepare schools, teachers, and districts for participation in the Year 3 PIE pilot study by establishing communication protocols, securing commitments, providing orientation and training, and delivering necessary tools, manuals, and technical support for successful pilot implementation. | Sixty fifth-grade teachers across 35 schools in 28 districts were recruited for the pilot study. Orientation meetings were held, support materials and manuals were developed and distributed, and the Kite system was set up for data entry and delivery. The Kite Service Desks staff were trained by ATLAS staff before the launch of the pilot study.<br><br>Several manuals and resources were developed for pilot study participants:<br><br>• *PIE Pilot Technical Specifications Manual*<br>• *PIE Pilot District Test Coordinator Manual*<br>• *PIE Pilot Test Administration Manual*<br>• *PIE Pilot Educator Portal User Guide*<br>• *PIE Pilot Practice Test Directions* |

| Objective | Objective outcomes | Project outcomes |
|---|---|---|
| | | The PIE assessment window opened on September 16, 2024, and closed on March 24, 2025. |

Goal 3 was to design an approach to evaluating technical adequacy, with a scoring model, theory of action, and validation plan for multiple uses, including future use as a statewide assessment. Overall, the intended outcomes of Goal 3 were met. Table 29 summarizes the objectives and outcomes relevant to Goal 3.

Table 29.

*Objectives and Project Outcomes for Goal 3*

| Objective | Objective outcomes | Project outcomes |
|---|---|---|
| **Objective 3.1**<br>Refine the draft theory of action to support a future, full-scale PIE system that could be used for summative purposes. | Develop a theory of action that clearly defines the intended purposes and uses of the PIE system. | Draft intended uses and a theory of action were developed, reviewed with input from Missouri DESE and the PAC, refined according to their feedback, and integrated into the final version. |
| **Objective 3.2**<br>Identify and conduct focused validity-evaluation studies for selected propositions in the theory of action | Develop propositions aligned with the PIE theory of action and plan the validity studies and evidence collection necessary to evaluate those propositions. | A complete set of draft propositions and corresponding evidence needs were developed, reviewed, and revised with PAC feedback, resulting in finalized plans for conducting focused validity studies in Year 3.<br><br>In Year 3, validity studies included interviews with six Missouri educators at two points (beginning and end of instructionally embedded window) to gather feedback on their use of the PIE system and how their practices changed over time. Additionally, end-of-pilot teacher focus groups with eight Missouri educators were conducted to collect insights on teacher perceptions and usage of the PIE system. Teacher and student surveys were administered to gather feedback on the PIE system and assessments. |

| Objective | Objective outcomes | Project outcomes |
|---|---|---|
| **Objective 3.3**<br>Apply and evaluate innovative scoring models to provide multiple metrics of student achievement | Develop and refine potential scoring models for the PIE system that provide multiple measures of student mastery and progress, support accountability and growth evaluations, and align with the PIE theory of action. | The hierarchical diagnostic classification scoring model (HDCM) was the statistical model used to establish the probability of mastery for each pathway level for PIE assessments during the pilot study instructionally embedded assessment window. The HDCM is a general latent-class model that uses a log-linear modeling framework to produce probabilities of class membership for multiple measured attributes that also restricts the latent classes that are permitted. Student mastery statuses for each pathway level were inferred using a Bayesian estimation method.<br><br>Several draft summative scoring models were presented to the PAC in October 2023 and revised according to their feedback. Findings from the pilot study data were later shared during the fall 2025 PAC meetings, specifically the technical evaluation of the scoring models (IRT, TDCM, Beta IRT, and Hybrid Beta IRT).<br><br>A synthesis of evidence relative to the claims in the PIE theory of action was developed. |
| **Objective 3.4**<br>Develop additional prototype concepts for potential future use in statewide assessment systems. | | Several concepts for the future-use, proof-of-concept model—including growth models, standard setting, and assessment designs—were discussed with the PAC in fall 2025. |

## Future Directions

Results and findings from the pilot study of the PIE system provide insights for designing and implementing an innovative instructionally embedded (through-year) assessment model. There are several areas for potential future research.

**Standards-Aligned, Research-Based Learning Pathways**

Future research on the use of standards-aligned, research-based learning pathways should focus on maximizing their utility for educators, enhancing teachers' content knowledge, and expanding their scope across grade levels.

Learning pathways are clearly useful to teachers by enabling them to establish more direct connections between their instructional materials and learning standards, fostering a better understanding of alignment to content standards in general. However, research should explore the most effective ways to provide additional support and resources to help both classroom teachers and curriculum support specialists see and leverage the connections between existing curricula and state standards. Future efforts might involve developing interactive tools, detailed mapping resources, or targeted professional development.

These learning pathways have shown promise in supporting teachers' content knowledge of math standards, which is particularly beneficial for teachers who hold general education degrees. Further investigation is needed into the types of additional training and learning opportunities that can best support teachers' understanding and effective use of the learning pathways in daily instruction. Research could assess the impact of sustained coaching models, collaborative learning communities, or micro-credentialing programs focused on the pathways.

Finally, the initial efficacy of this project was established using only fifth grade math standards. A significant direction for research is the development and study of vertically articulated learning pathways across grade levels 3-8 and in high school. Developing comprehensive pathways would be crucial to better meet the needs of all learners, providing educators with a map of the entire learning journey to accurately identify where each student is and plan instruction accordingly. This vertical view would also help address learning gaps and ensure smooth transitions between grades.

**Test Development**

Future research in test development should focus on processes that support efficient creation of robust and instructionally relevant item and task banks. Key directions could involve further optimizing item bank development, leveraging Artificial Intelligence (AI), and advancements in the integration of cognitive models with task design.

As demonstrated in the PIE project, evidence-centered design (ECD) provides a rigorous framework for linking claims about student knowledge and skills (the evidence) directly to the tasks that elicit that evidence. Future research is needed to explore methods for scaling up the application of ECD, perhaps through automated generation or parametrization of task shells, to rapidly populate item banks while maintaining alignment to content standards and desired psychometric properties.

Furthermore, use of AI to support the test development process is a rapidly evolving area of research and in some cases, practice. AI has the potential to transform multiple stages of the assessment life cycle. Research should investigate AI applications for:

- Automated item generation (AIG): Creating new test items, especially technology-enhanced items, from established task models.
- Automated item review: Using natural language processing (NLP) and machine learning (ML) to check items for bias, clarity, reading level, and alignment to standards before human review.
- Psychometric analysis: Developing AI algorithms that can predict item difficulty or discrimination parameters faster and with greater accuracy than traditional methods.

Finally, an area with many opportunities for future advancement involves using formal models of how students learn, often derived from cognitive science, coupled with ECD-based task models, to support the development of assessment tasks. Cognitive models (such as the learning pathways developed for the PIE project) can be used to guide the creation of assessment tasks that represent varying levels of cognitive complexity while remaining accurately aligned to content standards. While the PIE project leveraged the learning pathways in this way, future work could make the relationships between formal models of student learning, and cognitive process complexity of tasks designed to measure skills within the models, more explicit. This explicit integration, between cognitive domain and cognitive complexity of tasks, ensures that assessments measure higher-order thinking, problem-solving, and conceptual understanding in the context of formal learning models, which have the potential to provide more meaningful diagnostic information to students and teachers.

**Test Administration**

Future research and work on test administration designs that transform static, end-of-year testing into dynamic, instructionally integrated assessment systems should involve implementing adaptive technology and shifting to year-long, embedded assessment windows.

A critical future direction is the development and optimization of adaptive test engines, particularly those that use DCMs for scoring, to more efficiently measure individual student mastery. This approach is consistent with competency-based learning approaches, where assessment should occur when students are ready and target their current location within the learning pathways. However, research is needed to refine the algorithms of these engines to provide highly personalized item selection and scoring. This includes ensuring the assessment precisely reflects where students are in the learning journey and investigating how to seamlessly integrate these adaptive measurements with instructional data to provide teachers with real-time, actionable feedback.

**Instructionally Embedded Assessment System Implementation**

The broad-scale implementation of instructionally embedded assessment systems in general education presents a variety of questions and challenges to solve, moving beyond initial design to focus on scale, impact, and systemic change. The success of IEA depends on its acceptance and meaningful use by the people it is designed to serve. Continued research on the practical

implementation challenges, such as necessary teacher training and the feasibility of large-scale operational use (similar to what has been done in programs like Dynamic Learning Maps®), will be vital for broader adoption.

# References


Accessible Teaching, Learning, and Assessment Systems & Missouri Department of Elementary and Secondary Education [DESE]. (2024a). *PIE Pilot test administration manual*. https://pie.atlas4learning.org/sites/default/files/documents/resources/Pilot_Files/PIE_Pilot_Test_Administration_Manual.pdf

Accessible Teaching, Learning, and Assessment Systems & Missouri Department of Elementary and Secondary Education [DESE]. (2024b). *PIE Pilot educator portal user guide.* https://pie.atlas4learning.org/sites/default/files/documents/resources/Pilot_Files/PIE_Pilot_Educator_Portal_User_Guide.pdf

Accessible Teaching, Learning, and Assessment Systems. (2025a). *Pathways for Instructionally Embedded Assessment: PIE assessment design and development technical report*. University of Kansas; Accessible Teaching, Learning, and Assessment Systems. https://pie.atlas4learning.org/sites/default/files/documents/resources/PIE_Assessment_Design_Development_Technical_Report.pdf

Accessible Teaching, Learning, and Assessment Systems. (2025b). *Educator perspectives on the PIE assessment system: Evidence from the PIE pilot study.* University of Kansas; Accessible Teaching, Learning, and Assessment Systems.

Accessible Teaching, Learning, and Assessment Systems. (2025c). *Evaluating Pathways for Instructionally Embedded Assessment (PIE) assessments: Evidence from cognitive labs.* University of Kansas; Accessible Teaching, Learning, and Assessment Systems. https://pie.atlas4learning.org/sites/default/files/documents/resources/PIE_Evidence_From_Cognitive_Labs.pdf

Accessible Teaching, Learning, and Assessment Systems. (2025d). *PIE as a proof of concept for future summative use: Evidence from a pilot study.* University of Kansas; Accessible Teaching, Learning, and Assessment Systems.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*(4), 541–561. https://doi.org/10.1007/bf02296195

Bennett, R. E. (2018). Educational assessment: What to watch in a rapidly changing world. *Educational Measurement: Issues and Practice*, *37*(4), 7–15. https://doi.org/10.1111/emip.12231

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Choi, K. M., Hwang, J., Jensen, J., & Hong, D. S. (2022). Teachers' use of assessment data for instructional decision making. *International journal of mathematical education in science and technology*, *53*(4), 1010–1017. https://doi.org/10.1080/0020739x.2021.1880653

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://psycnet.apa.org/doi/10.1037/0033-2909.112.1.155

Design Council. (2005). *A study of the design process—The double diamond*. https://www.designcouncil.org.uk/fileadmin/uploads/dc/Documents/ElevenLessons_Design_Council%2520%25282%2529.pdf

Dynamic Learning Maps Consortium. (2022). *2021–2022 Technical manual—Instructionally Embedded model*. University of Kansas; Accessible Teaching, Learning, and Assessment Systems. https://2022-ie-techmanual.dynamiclearningmaps.org/

Every Student Succeeds Act, 20 U.S.C. § 6301. (2015).

Guskey, T. R. (2022). *Implementing mastery learning*. Corwin Press.

Henson, R. A., Templin, J., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191–210. https://doi.org/10.1007/s11336-008-9089-5

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70. http://www.jstor.org/stable/4615733

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329–349. https://doi.org/10.1207/s15324818ame1404_2

Johnson, M. S., & Sinharay, S. (2018). Measures of agreement to assess attribute-level classification accuracy and consistency for cognitive diagnostic assessments. *Journal of Educational Measurement*, *55*(4), 635–664. https://doi.org/10.1111/jedm.12196

Karvonen, M., Ruhter, L., Shipman, M., Swinburne Romine, R., & Tiemann, G. (2020). *Designing, developing, and evaluating innovative science assessments: Evidence from the I-SMART project* [Technical report]. University of Kansas; Accessible Teaching, Learning, and Assessment Systems. https://ismart.works/sites/default/files/documents/Publications/I-SMART_Goal2_TestDev_TechnicalReport_FINAL.pdf

Kim, E. M., Nash, B., & Swinburne Romine, R. (2024). *Pathways for instructionally embedded assessment (PIE): Developing learning pathways for the PIE assessment system*. Lawrence, KS: University of Kansas; Accessible Teaching, Learning, and Assessment Systems. https://pie.atlas4learning.org/sites/default/files/documents/resources/Developing_Learning_Pathways_for_the_PIE_Assessment_System.pdf

Marion, S. F. (2018). The opportunities and challenges of a systems approach to assessment. *Educational Measurement: Issues and Practice*, *37*(1), 45–48. https://doi.org/10.1111/emip.12193

Martin, C., L., Mraz, M., & Polly, D. (2022). Examining elementary school teachers' perceptions of and use of formative assessment in mathematics. *International Electronic Journal of Elementary Education*, *14*(3), 417-425. https://doi.org/10.26822/iejee.2022.253

Nash, B., Majerus, M., Bates, S., & Gane, B. (2024, April 11–14). *Designing assessments to support instruction* [Conference presentation]. Annual meeting of the National Conference on Measurement in Education, Philadelphia, PA, United States. https://pie.atlas4learning.org/sites/default/files/documents/resources/Designing%20Assessments%20to%20Support%20Instruction%20NCME%202024.pdf

Nash, B., Majerus, M., & Tegart, T. (2025, April 23–26). *Designing professional learning to support teachers' use of an innovative assessment system intended to inform classroom instruction.* [Conference presentation]. Annual meeting of the National Conference on Student Assessment, Denver, CO, United States. https://pie.atlas4learning.org/sites/default/files/documents/resources/Designing_Professional_Learning_for_PIE_Assessments_NCSA_2025.pdf

Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). *The role of interim assessments in a comprehensive assessment system: A policy brief.* The Aspen Institute. https://www.achieve.org/files/TheRoleofInterimAssessments.pdf

Speckesser, S., Runge, J., Foliano, F., Bursnall, M., Hudson-Sharp, N., Rolfe, H., & Anders, J. (2018). *Embedding formative assessment: Evaluation report and executive summary*. Education Endowment Foundation. https://dera.ioe.ac.uk/32012/1/EFA_evaluation_report.pdf

Stan Development Team. (2022). *RStan: The R interface to Stan*. https://mc-stan.org/

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*(2), 317–339. https://doi.org/10.1007/s11336-013-9362-0

Thompson, W. J. (2019). *Bayesian psychometrics for diagnostic assessments: A proof of concept* (Research report No. 19–01). University of Kansas; Accessible Teaching, Learning, and Assessment Systems. https://dynamiclearningmaps.org/sites/default/files/documents/publication/Bayes_Proof_Concept_2019.pdf

Thompson, W. J. (2023). measr: Bayesian psychometric measurement using Stan. *Journal of Open Source Software, 8*(91), Article 5742. https://doi.org/10.21105/joss.05742

Varier, D., Powell, M. G., Dodman, S., Ives, S. T., DeMulder, E., & View, J. L. (2024). Use and usefulness of assessments to inform instruction: Developing a K–12 classroom teacher assessment practice measure. *Educational Assessment*, *29*(1), 36–57. https://doi.org/10.1080/10627197.2024.2316907

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, *16*(2), 667–718. https://doi.org/10.1214/20-BA1221

von Davier, M., & Haberman, S. J. (2014). Hierarchical diagnostic classification models morphing into unidimensional 'diagnostic' classification models—A commentary. *Psychometrika*, *79*(2), 340–346. https://doi.org/10.1007/s11336-013-9363-z

Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF* [Working paper]. University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science.

# Appendix A: Pilot Study Activities (Classroom Teacher)

| Dates (Fall 2024–Spring 2025) | Activity | Estimated time to complete |
|---|---|---|
| 8/15; 8/16 | Attend an online orientation meeting (via Zoom) | 1 hour |
| 8/19–8/23 | Visit (and bookmark) the PIE pilot study web page at pie.atlas4learning.org/pilot<br>• Click on the Kite resource page (at the bottom of the page).<br>• Log in to Educator Portal.<br>• Complete the security agreement.<br>• Navigate to the training tab and launch the pilot training course.<br>Begin teacher training course.<br>• Complete modules 1 and 2.<br>• Review job aids and resources. | 1 hour |
| 8/26–9/13 | Create content groups.<br>Develop/update your classroom instructional plans as needed (i.e., according to the scope and sequencing of your content group selections). | ~ 2 hours |
| 9/3–9/6 | Continue teacher training course.<br>• Complete modules 3 and 4.<br>• Review job aids, resources, and the *PIE Pilot Test Administration Manual* (ATLAS & DESE, 2024a).<br>Administer the technology practice items with your students (9/3–9/13). | 1 hour<br>~ 25 minutes |
| 9/9–9/13 | Continue teacher training course.<br>• Complete modules 5 and 6.<br>• Review job aids, resources, and user manuals as needed. | 1 hour |
| 9/16–2/19 | Complete assessment and instruction cycles according to your developed plan.<br>• Each assessment and instruction cycle may take 2–4 weeks.<br>• Within each cycle, three short diagnostic assessments will be administered: baseline, midway, and end of unit. | 18 calendar weeks |

| Dates (Fall 2024–Spring 2025) | Activity | Estimated time to complete |
|---|---|---|
| | • Each assessment consists of 6–20 items depending on the size of the content group.<br>• It is anticipated that each assessment will take about 10–25 minutes to complete, depending on the size of the content group.<br>• Assessment results will be immediately available for teachers to view and use to inform instruction. | |
| 2/1–2/21 | Complete a survey about your experience using PIE. | 30 minutes |
| 2/1–2/21 | Administer a short survey to your students about their experiences with the PIE assessments. | 5 minutes |
| 2/24–3/21 | Administer the final assessment. | ~ 1.5 hours |