PIE

Pathways for Instructionally Embedded Assessment

PIE as a Proof-of-Concept for Future Summative Use: Evidence From a Pilot Study

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. The following project members made significant writing contributions to this report: Auburn Jimenez, Brooke Nash, and Jake Thompson.

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.

Table of Contents

1 Background

Through-year assessments have received heightened interest in recent years as a potential solution to pushback against the traditional high-stakes, end-of-year test that most states use to evaluate student achievement. Stakeholders and advocacy groups have rightly noted that, while traditional assessment models provide only a single snapshot that has little instructional utility, through-year assessments can provide more frequent and more fine-grained feedback to support instructional decision-making and student learning. However, even with a renewed focus on assessment for the purpose of learning, we still must recognize the need for a summative score, for both succinctly describing achievement to parents and students and for use in state accountability models.

A key consideration for any summative score is the claim we wish to make about student achievement. This is especially true for through-year assessments, which can be designed in many ways. Traditional end-of-year assessments also have a variety of designs, but they all are administered toward the end of the academic year to make a claim about student achievement *at the end of the year*. In contrast, through-year assessments collect data throughout the year at different intervals, depending on the design of the system. Therefore, through-year assessments are making a claim about student achievement *by the end of the year*, but specific claims about the timing of achievement and the design of the system must inform each other.

In this report, we describe the evaluation of summative scoring models for a through-year assessment that is grounded in the range of claims that may be made by a through-year assessment system. We begin with a brief discussion of example through-year assessment designs and the claims that each design may support. We then discuss a wide range of summative scoring models that could support different claims. Throughout, we use the Pathways for Instructionally Embedded Assessment (PIE) project as a demonstrative example. PIE is a proof-of-concept through-year assessment developed in a partnership between Accessible Teaching, Learning, and Assessment Systems (ATLAS) at the University of Kansas and the Missouri Department of Elementary and Secondary Education (DESE). This collaboration was funded through the U.S. Department of Education's Competitive Grants for State Assessments program. We apply each summative model to data collected from a pilot study of the PIE system to demonstrate how different assessment design choices and desired claims affect the feasibility of different modeling approaches. Finally, we discuss how other assessment system components (e.g., standard setting, growth models) may also influence the choice of a summative scoring model.

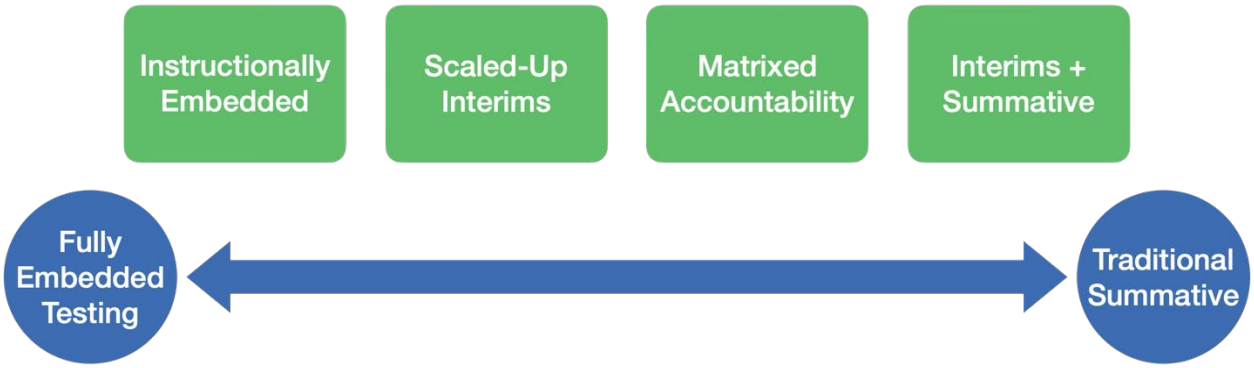
1.1 Through-Year Assessment Models

Broadly, through-year assessments use evidence gathered throughout the year to produce a result at the end of the year that may be used for summative reporting purposes (Dadey et al., 2023; Gong, 2025). Within this definition, many designs would support different claims related to the specific timing of student achievement. A full discussion of possible

through-year designs is beyond the scope of this report. Rather, we focus on a few broad categories of designs that represent the spectrum of possible designs, ranging from fully embedded testing to the traditional spring-only summative (Figure 1). Specifically, here we focus on instructionally embedded, scaled-up interim assessments, and matrixed accountability designs.

Figure 1

Possible Designs of Through-Year Assessment System



1.2 Overview of Through-Year Designs

Instructionally embedded assessment systems are characterized by short but frequent assessments that are embedded into teachers' typical instructional units. Teachers may flexibly choose which standards to assess and when, according to local curriculum and pacing guides. There is no final or end-of-year assessment—results are based on the assessments completed during the year. These assessments are built on research-based models of student learning and cognition to support instructional decision-making. Results from the assessments are immediately available, asset oriented, and understandable to multiple stakeholder groups (e.g., teachers, parents, students). Additionally, the results are provided at a grain size that is consistent with the intended use of informing instructional decisions and thus also demonstrate connections to precursor and successor skills so that students have clear next steps for learning after each assessment. At the end of the year, the fine-grained results provided throughout the year are aggregated (e.g., into an overall achievement level) to support summative reporting. Examples of this approach include the Dynamic Learning Maps® (DLM®) alternate assessment (Dynamic Learning Maps [DLM] Consortium, 2022) and the PIE project (Accessible Teaching, Learning, and Assessment Systems [ATLAS], 2025).

Scaled-up interim systems consist of assessments that are administered at specific times during the year. For example, a system may include fall, winter, and spring assessments; however, the number and frequency of assessments may differ significantly across variations of this design. Additionally, there may or may not be an end-of-year component. If an end-of-year component is included, it should serve a specific purpose (e.g., to assess previously untested standards, optional time to demonstrate mastery if additional instruction occurred after an interim was administered). These assessments are not

typically based on models of student learning, but rather are divisions of an assessment blueprint. In some models, each interim is a mini summative, meaning that the full blueprint is covered on each assessment but with a reduced number of items. Alternatively, a model may segment the blueprint so that each assessment covers a subset of standards and all standards are covered across the set of assessments. Instructional utility of scaled interims is necessarily driven by the frequency and grain size of feedback and flexibility in content assignments. An assessment system consisting of three interims in which the content is determined by district staff for the purpose of monitoring programmatic or curricular effectiveness may be less instructionally informative than a system with five to seven interims in which the content assessed within each window is chosen by each teacher or by school or district staff in alignment with local pacing guides. Examples of this interim approach include Innovative Assessment Demonstration Authority (IADA) programs in Louisiana (Munyan-Penney & Barone, 2020) and Missouri (DESE, 2025), as well as the Montana Aligned to Standards Through-Year Assessment (MAST; Marion et al., 2022).

We define *matrixed accountability designs* as designs that are characterized by a distinction between student-level and school- or district-level claims about student achievement. In these designs, student-level claims are focused on achievement *by the end of the year*, and therefore assessments may follow an instructionally embedded or scaled-up interim design as described in the preceding paragraph However, a state may wish to measure school or district achievement *at the end of the year* (e.g., to inform programmatic or curricular decisions for the following school year). To support claims for *at the end of the year*, this model would include an end-of-year assessment. However, the goal is not to measure student-level performance. Rather, student-level results continue to be driven by data collected during the year. Because we are not trying to support student-level reporting at the end of the year, the end-of-year assessment can use matrix sampling to reduce the spring assessment burden. Therefore, no student is assessed on all standards, but across each school or district, all standards are assessed. However, within matrixed accountability designs, it is possible to use the end-of-year assessment for multiple purposes. For example, the assessment may be matrix sampled but prioritize untested standards for students who did not cover the full blueprint during the year. The matrixed accountability design is similar to the approach taken by the National Assessment of Educational Progress (NAEP; Thurgood et al., 2003), where assessment results are used only for aggregated reporting.

1.3 Claims Supported by Different Through-Year Designs

Because of the timing of assessments in a through-year assessment system, designs can support different claims of student achievement. However, all claims fall under the larger umbrella claim that should be supported by all through-year systems: “Summative results accurately reflect achievement of grade-level academic standards by the end of the year.”

Instructionally embedded assessments are flexibly administered in alignment with each teacher’s normal instructional cycle. Therefore, an instructionally embedded design could

support a claim about achievement *at the time of instruction*. In contrast, scaled-up interim designs could support a claim about achievement *at specific intervals*, where the intervals are determined by the specific interim design. Finally, the matrixed accountability design would support a student-level claim (depending on the design of the through-year component) as well as a school- or district-level claim about target achievement *at the end of the year*.

Table 1 summarizes these claims, the assessments administered during the year and at the end of the year, and the data used to generate summative scores.

Table 1

Assessment Designs and Summative Scoring Data, by Claim

Claim	Assessment during the year	Final assessment	Summative scoring
Students demonstrate mastery at the time of instruction throughout the year.	Instructionally embedded; flexible, instructionally useful	N/A	During-year data
Students demonstrate mastery at specific intervals during the year.	Scaled interims; flexibility and utility driven by system design	Optional; scaled down blueprint to serve a specific purpose	During-year data + final data
Schools/districts are on target at the end of the year.	Instructionally embedded or scaled interims	Reduced blueprint to serve accountability needs only	Students = during-year data; School/district = final data

2 The PIE System

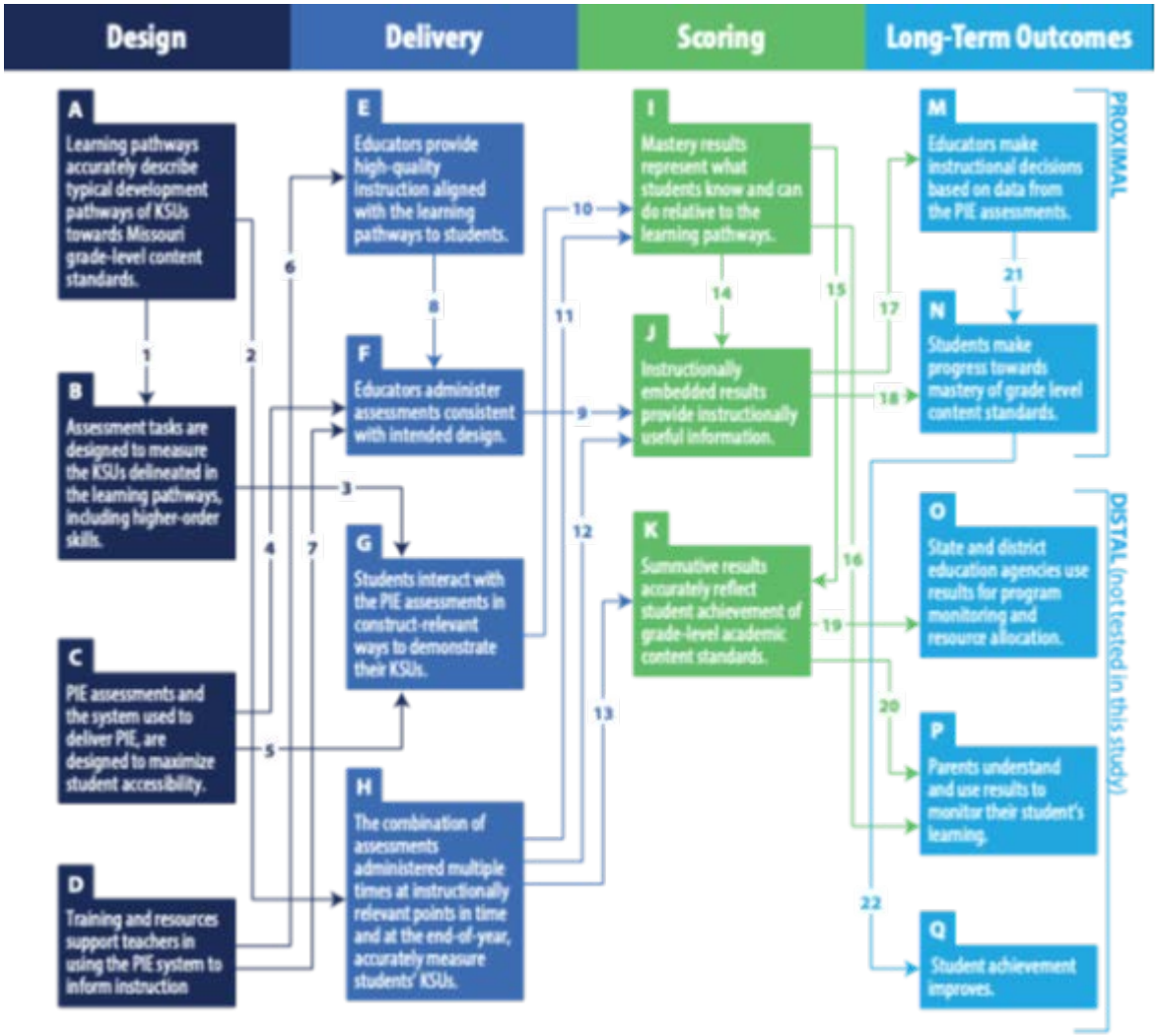
The PIE system is a proof-of-concept, instructionally embedded model of assessment in general education fifth grade mathematics. The assessments are built on learning pathways that were developed for 25 Missouri priority mathematics standards. Each pathway consists of three levels: the target (referred to as Level 3, aligned to the grade-level standard) and two precursor skills (Level 2 and Level 1). The levels make up a linear progression: mastery of Level 1 is a prerequisite for Level 2, which is in turn a prerequisite for Level 3. For a complete description of the development of the learning pathways, see Kim et al. (2024). During the year, teachers can select the standards for which they are

providing instruction and choose when students are ready to receive assessments on the relevant skills.

The PIE project used an argument-based approach to assessment validity, which included a three-tiered approach. Tier 1 specified a theory of action defining claims in the validity argument that support the goals of the system (Figure 2). Tier 2 specified an interpretive argument that defined the evidence that should be evaluated to support each claim. Finally, Tier 3 consisted of the actual validity studies conducted to evaluate each claim. A system pilot study was administered during the 2024–2025 school year to collect evidence to evaluate the claims in the theory of action.

Figure 2

PIE Theory of Action



Note. KSUs = knowledge, skills, and understandings.

Most relevant to this report is Claim K: "Summative results accurately reflect student achievement of grade-level academic content standards." As PIE is a proof-of-concept project, this claim was kept intentionally vague regarding timing so that we could evaluate a variety of summative designs and scoring models (e.g., inclusion vs. exclusion of final assessment data). Those analyses and how our findings may inform other through-year designs are the focus of this report. For a summary of evidence supporting other claims in the PIE theory of action, including the use of PIE assessments to inform instruction, we direct readers to the *PIE Pilot Study: Design and Administration Evidence* technical report (ATLAS, 2025).

2.1 Design and Delivery of Pilot Study

The PIE pilot study consisted of two windows for assessment delivery. The first was the instructionally embedded window, which lasted from September 2024 through February 2025. The second was the final window, which was open from February 2025 through March 2025. During the instructionally embedded window, educators selected subsets of the 25 content standards, guided students through concepts embedded in a three-level learning pathway for each of the selected standards, and assessed students on concepts over three testing occasions: baseline, midway, and end-of-unit.

At the start of an instructional cycle, a baseline assessment is administered to evaluate proficiency in Level 1 skills. Instruction then occurs between the baseline and midway assessments, allowing educators to address identified needs before a midway assessment is administered to assess proficiency in Level 2 skills. After midway results are reviewed, a second phase of instruction occurs before the end-of-unit assessment, which measures Level 3 skills. Students who had not yet demonstrated mastery of Level 1 skills were retested at the midway assessment, and those who had not yet demonstrated mastery of Level 2 skills were retested at the end-of-unit assessment, ensuring multiple opportunities to show growth within the instructional cycle. The final component consisted of a 50-item fixed-form assessment of Level 3 skills from all 25 standards (two items per standard). Across both windows, the item responses consisted of a mix of dichotomous items (e.g., correct or incorrect) and polytomous items (e.g., ordered response categories reflecting increasing levels of outcomes).

2.2 Participation

This pilot study was approved by the University of Kansas Center for Research Institutional Review Board (ID: STUDY001499737) and conducted in full compliance with ethical guidelines, including approved procedures for recruitment, consent, and data handling.

For a student to be eligible for inclusion in the present study, that student must have fully tested on at least one content standard within the instructionally embedded administration window. In total, 1,572 fifth grade students in Missouri met this criterion and participated in the 2024–2025 PIE pilot study, representing 55 teachers across 32 schools and 28 districts. The student population was evenly split by gender and was predominantly White and non-Hispanic (Table 2).

Table 2

*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	[a]	[a]
Native Hawaiian or Pacific Islander	[a]	[a]
Alaska Native	[a]	[a]
Hispanic 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

Note. [a] Data suppressed because sample size was <10.

During the pilot administration, student participation varied across both the instructionally embedded and final assessment components. During the instructionally embedded window, participation levels differed widely, with 28% of students completing one to five of the expected 25 standards, 35% completing six to 10 standards, 20% completing 11 to15 standards, 9% completing 16 to 20 standards, and 9% completing 21 to 25 standards. On average, students completed assessments for 2.61 of the three levels within each standard. During the final component, 1,187 students completed all 25 standards via a single, fixed form assessment; 385 students who participated in the instructionally embedded component did not complete the final component. For a complete description of pilot study participation and patterns of usage of the system during the pilot, see the *PIE Pilot Study: Design and Administration Evidence* technical report (ATLAS, 2025).

3 Summative Scoring for Through-Year Assessments

Generating a summative score from a through-year assessment presents unique challenges. The timing and content coverage of assessments in a through-year design may vary across testing occasions, creating differences in measurement conditions that must be addressed in the scoring model. For example, there may be instances when some students are given extra items (e.g., in the case of nonmastery at previous time points as in PIE), creating intentional patterns of missing data. Additionally, students' knowledge and

skills are intended to change between testing occasions, introducing dependencies in the data that violate assumptions of independence and stability of the latent construct, which may require the use of a longitudinal model. Finally, like any other assessment program, summative results must be grounded in the theory of action to meet the intended claims. Furthermore, results must preserve the statistical properties necessary for reliable, valid, and interpretable reporting, while still accounting for the other unique aspects of a through-year design.

Together, these challenges underscore the need for multiple modeling frameworks—such as item response theory (IRT), diagnostic classification models (DCMs), and hybrid DCM/IRT approaches—that can integrate diverse data sources, account for measurement error, model changes in student ability that occur across the full assessment window, and meet the specific claims of the assessment system.

In this section, we describe the various summative scoring models used for the PIE project to evaluate the psychometric properties related to different design decisions. Specifically, we examined IRT, DCM, and hybrid models. For each model, we considered variants that did or did not include data from the final component. We also considered variants that included items aligned to all three pathway levels and variants that included only items aligned to Levels 2 and 3, which were aligned to direct precursor skills and the targeted standard, respectively. Note that in this section we explore different modeling options to demonstrate how this process may develop for different claims that a through-year system may intend to make. At the end of this section, we return to the PIE theory of action to evaluate how each model would or would not support claims of the PIE project.

We also note that the model chosen for generating summative results is not purely a function of psychometric properties and intended claims. The choice is likely also influenced by policy considerations, such as standard setting methods, growth models, and other accountability indicators. We discuss the role of these factors in Section 4.

3.1 Item Response Theory (IRT)

In the present study, we employed well-established IRT models to accommodate the item formats used in the pilot administration. We selected two models: the two-parameter logistic (2PL; Birnbaum, 1968) model for dichotomous items and its polytomous extension, the graded response model (GRM; Samejima, 1969), for ordered categorical items. The 2PL model supports estimation of item-level slope parameters, which indicate how well items discriminate among students with varying ability levels. The GRM extends the 2PL framework to accommodate items with multiple ordered response categories.

3.1.1 Data Conditions

To evaluate measurement quality under different data conditions, three IRT models were estimated, each defined by the assessment design and the item types included. The *IRT-IE* model was based solely on instructionally embedded administrations at baseline, midway, and end-of-unit, covering items aligned to Levels 1–3. The *IRT-Final* model drew exclusively

on final items, which targeted Level 3 skills. The *IRT-Combined* model incorporated both instructionally embedded and final items (Table 3).

Table 3

Item Response Theory (IRT) Models and Data Conditions

Model	IRT scale score data	Levels included
IRT-IE	Baseline, midway, end-of-unit	Level 1, level 2, level 3
IRT-Final	Final	Level 3
IRT-Combined	Baseline, midway, end-of-unit, final	Level 1, level 2, level 3

3.1.2 Model Definitions

In an IRT framework, each item directly measures the overall summative latent trait. For the PIE models, we estimated a 2PL IRT model for dichotomous items (e.g., multiple-choice single-select, multiple-choice multiple-select) and a GRM for polytomous items (e.g., composite, constructed response). For details on how these models are defined, see Birnbaum (1968) and Samejima (1969), respectively. All models were estimated using a Bayesian framework, which offers more flexible and more robust options for evaluating model fit than traditional maximum likelihood approaches do (Thompson, 2019). Estimation was carried out using Markov chain Monte Carlo procedures implemented in the *rstan* software package (Stan Development Team, 2025), an interface to the *Stan* probabilistic programming language (Carpenter et al., 2017).

Prior distributions were specified consistently across models. Intercept parameters were assigned a normal prior with mean 0 and variance 2, providing weakly informative but unbounded support. Main-effect parameters were constrained to be positive to ensure monotonicity and were assigned lognormal priors with mean 0 and variance 1.

Convergence of the estimation procedure was assessed following the recommendations of Vehtari et al. (2021), requiring all $\hat{R}$ values to be below 1.01 and effective sample sizes to exceed 400 to ensure that the posterior parameter space was adequately explored across the four chains. After confirming convergence, posterior means were used to summarize parameter estimates.

3.1.3 Model Evaluation

Absolute model fit was evaluated using posterior predictive model checks (PPMCs), a simulation-based approach (e.g., Thompson, 2019). At each iteration of the Bayesian sampling algorithm, parameter estimates were used to generate a replicated dataset per model for comparison with the observed data. This procedure produced 8,000 replicated datasets that mirror the structure of the observed responses. By comparing summaries of the replicated and observed data, we assessed the degree to which each model was able to reproduce key patterns in student performance.

Posterior predictive *p*-values (*ppp*) were used to evaluate the absolute fit of the IRT models across the three data sources. The IRT-IE model captured patterns in the observed data, indicated by a *ppp* value of .261, suggesting adequate fit. In contrast, the IRT-Final model poorly captured patterns in the observed data, indicated by a *ppp* value of .005, suggesting that final assessment response data were not sufficient to support adequate fit. The IRT-Combined model, which accumulates data from the instructionally embedded and Final windows, demonstrated adequate fit with a *ppp* value of .272. This pattern may reflect key differences in the delivery design of the assessment windows. The instructionally embedded window encompassed all 25 standards and three pathway levels per standard, each level measured by three to five items. Student participation was also higher, providing more evidence for model estimation. In contrast, the final assessment window tested only the target (Level 3) skills and offered only two items per standard. Participation was lower as well, which reduced the amount of information available to estimate the necessary model parameters. These factors likely contributed to the superior fit of the IRT-IE model and the limited added value of final data when combined with the instructionally embedded data (Table 4).

Table 4

Item Response Theory (IRT) Model Fit Evaluation

Model	Posterior predictive *p*-value	Fit evaluation
IRT-IE	.261	Adequate
IRT-Final	.005	Poor
IRT-Combined	.272	Adequate

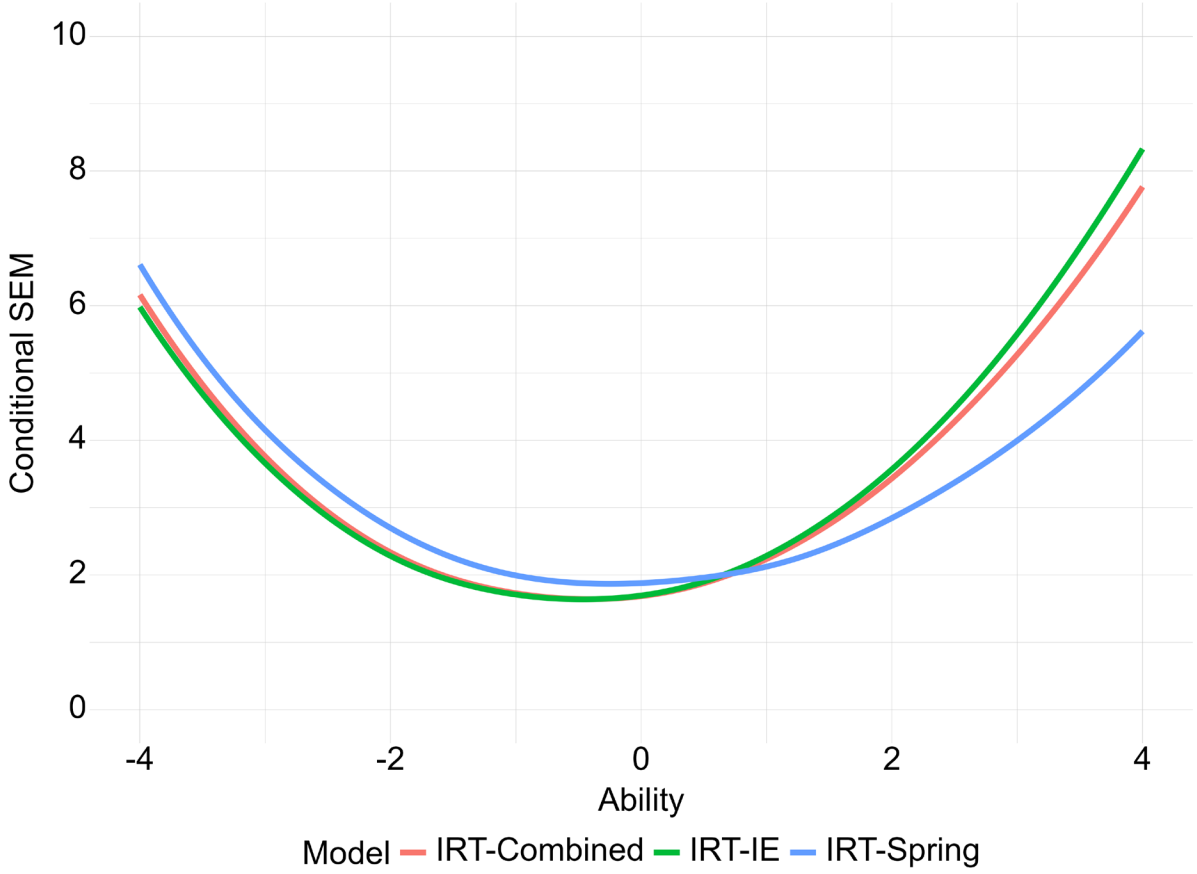
For the two-parameter logistic (2PL) and the graded response model (GRM), the reliability of ability estimates is evaluated through the test information. Test information represents how much information a test provides about a student's latent ability, and it varies across the ability scale. The test information function (TIF) summarizes this relationship, highlighting the regions of the scale where the test is most informative. Reliability can also be expressed in terms of error. The conditional standard error of measurement (CSEM) at a given ability level is the inverse of the TIF. Together, these concepts establish the foundation for evaluating the reliability of IRT models. However, it is important to note that models with larger item pools appear more reliable because test information and CSEMs estimate scale alongside the number of items, regardless of whether those additional items provide meaningful information beyond what is already captured. Therefore, it is useful to examine measurement efficiency, where precision is scaled by test length. This approach allows for fairer comparisons of reliability across models.

Reliability patterns were consistent across the IRT models that were evaluated (Figure 3). For each model, test information peaked and CSEMs were minimized near the center of the ability scale, as is normally expected. The IRT-IE model provided the greatest amount of test information and the lowest CSEM, indicating that this model produced score estimates with the strongest precision. In contrast, the IRT-Final model demonstrated the

weakest reliability, providing less information about examinees' ability across the scale. This result is not surprising in this context, given how many fewer items were included on the final assessment relative to the number of items students completed during instructionally embedded testing. The combined model provides similar amounts of information about examinees to the IRT-IE model. These findings suggest that for the PIE design, including the final component in a summative IRT model did not meaningfully affect the reliability of the student results.

Figure 3

Conditional SEM for IRT Models, Scaled by Test Length



3.1.4 Summary of IRT Models in Through-Year Designs

Results from the PIE pilot study show that the IRT model using only instructionally embedded data yielded adequate model fit and strong measurement precision for producing summative scale scores. In contrast, the model that was estimated solely from final assessment data showed limited fit and lower reliability. Combining instructionally embedded and final assessment data added little information beyond what was already provided by the instructionally embedded window. However, these findings are likely due to the test design. Whereas the instructionally embedded window included about five

items per level per standard (approximately 375 total items), the final assessment comprised only 50 Level 3 items. Thus, that component added relatively little information on its own.

3.2 Diagnostic Classification Models (DCMs)

While IRT models focus on placing students along a continuous ability scale, diagnostic classification models (DCMs; Rupp et al., 2010) provide fine-grained, actionable information about students' mastery of specific skills. Rather than estimating a single latent trait, DCMs classify students into mastery profiles that reflect which skills they have or have not yet acquired. These mastery indicators can be used to guide instruction, providing teachers with evidence to target areas of student need.

Transition diagnostic classification models (TDCMs; Madison & Bradshaw, 2018) capture changes in mastery status across multiple time points. By modeling how students transition between mastery and nonmastery states over time, TDCMs are well suited for through-year assessment designs, where the goal is to monitor learning progress and provide instructional feedback throughout the academic year.

3.2.1 Data Conditions

Three TDCM variants are evaluated as a proof-of-concept for future through-year assessment. Two TDCM variants—TDCM-IE and TDCM-Combined—are designed to evaluate how effectively student learning can be tracked over time across a progression of increasingly complex skills within defined learning pathways. The TDCM-IE model relies solely on instructionally embedded assessment to test whether reliable evidence of student mastery can be obtained without a final component. This is a three-time-point model, where the time points correspond to the baseline, midway, and end-of-unit assessments for each standard. The TDCM-Combined model incorporates instructionally embedded and final assessments (i.e., adding a fourth time point) to determine whether adding the final component improves overall adequacy of the model and the reliability of student scores for tracking learning by the end of the year. We also evaluated a two-time-point TDCM (TDCM-L3) that focused specifically on Level 3 skills, estimating changes in mastery status between the end-of-unit and final assessments. While the TDCM-Combined model includes Level 3 growth as part of its broader framework, the L3 variant includes only the levels that are directly aligned to the grade-level standard. This simpler model also can be used to evaluate changes between the instructionally embedded and final assessment components (e.g., growth after additional instruction, regression after time has passed; Table 5).

Table 5

DCM Models and Data Conditions

Model	DCM mastery probability data	Levels included
TDCM-IE	Baseline, midway, end-of-unit	Level 1, level 2, level 3
TDCM-Combined	Baseline, midway, end-of-unit, final	Level 1, level 2, level 3
TDCM-L3	End-of-unit, final	Level 3

3.2.2 Model Definitions

Each PIE learning pathway was calibrated using a TDCM, defined as a longitudinal extension of the log-linear cognitive diagnosis model (LCDM; Henson et al., 2009). Because each of the three pathway levels was represented at every time point, a learning pathway measured across T time points contains $3 \times T$ binary attributes. Each assessment item measured a single pathway level, creating a simple Q-matrix structure. The model estimates mastery probabilities for each attribute and time point, as well as transition probabilities describing changes in mastery from one time point to the next. Student results are reported as posterior probabilities of mastery, with values between 0 and 1 for each pathway level and time point.

To reflect the sequential nature of the PIE learning pathways, hierarchical constraints were imposed so that a student could not be classified as having mastered a higher level without first mastering all preceding levels (Templin & Bradshaw, 2014). This restriction reduces the number of permissible mastery profiles and aligns the model with the intended learning progression. All models were estimated using maximum likelihood methods implemented in the *Stan* probabilistic programming language via the *rstan* package (Carpenter et al., 2017; Stan Development Team, 2025). For technical details on model specification and scoring, see Madison and Bradshaw (2018).

3.2.3 Model Evaluation

Within DCM frameworks, the reliability of mastery estimates is evaluated through classification accuracy and classification consistency. *Classification accuracy* describes how well the model's mastery classifications reflect students' true skill levels, while *classification consistency* indicates how stable those classifications would remain if students were tested again under equivalent conditions. Both statistics can be computed using estimates from the fitted model and are reported for each pathway level and time point. Together these measures provide complimentary evidence of reliability and establish the confidence with which pathway-level mastery classifications can be interpreted in PIE. For additional technical details, see Johnson and Sinharay (2018). Table 6 presents guidelines for interpreting the two reliability measures.

Table 6

Interpretive Guidelines: Classification Accuracy and Classification Consistency

Category	Classification accuracy	Classification consistency
Excellent	≥.99	1.00
Very good	.95–.98	.99
Good	.89–.94	.86–.98
Fair	.83–.88	.70–.85
Poor	.55–.82	<.70
Weak	<.55	—

The reporting of results from the reliability analysis reflects decisions in the test design. Levels 2 and 3 were not assessed at baseline and Level 3 was not assessed at midway, so we do not report these pathway levels at these time points in our findings. For the TDCM-IE models, we evaluated 150 pathway level and time point combinations for classification accuracy and classification consistency: 25 × 3 = 75 pathway levels and time points for Level 1, 25 × 2 = 50 pathway levels and time points for Level 2 and 25 × 1 = 25 pathway levels and time points for Level 3. For the TDCM-Combined models that extend to a fourth time point, we evaluated 225 pathway level and time point combinations for classification accuracy and classification consistency: 25 × 4 = 100 pathway levels and time points for Level 1 25 × 3 = 75 pathway levels and time points for Level 2 and 25 × 2 = 50 pathway levels and time points for Level 3. Last, for the TDCM-L3 models that focus on Level 3 skills only, we evaluated 25 × 2 = 50 pathway level and time point combinations for classification accuracy and classification consistency.

3.2.3.1 TDCM-IE classification accuracy and consistency

Figure 4 summarizes the percentage of pathway levels within each classification accuracy category for the TDCM-IE model, aggregating results for all time points across standards. Level 1 was shown to be strong at baseline and midway, demonstrating *good* and *very good* levels of accuracy, with a slight decline at end-of-unit. Level 2, which was reported only at midway and end-of-unit, consistently demonstrated *good* and *very good* accuracy. Last, Level 3, which was reported only for end-of-unit, demonstrated mostly *fair* and *good* levels of accuracy. Differences in classification accuracy for the same level (e.g., Level 1) across time points reflect the assessment design and how information is incorporated into the models. For example, Level 1 is directly measured at baseline and midway, but not at end-of-unit, where its estimates are inferred indirectly through structural relationships with higher pathway levels (i.e., Level 2, Level 3). In addition, Level 1 is retested at midway and Level 2 is retested at end-of-unit, but only for students who did not demonstrate mastery at the prior occasion. Thus, the Level classifications are updated at each time point with new data, so accuracy for a Level may change over time. Overall, 141 pathway levels and time points (94%) demonstrated at least *fair* classification accuracy.

Figure 4

Estimated Classification Accuracy Metrics for TDCM-IE Models

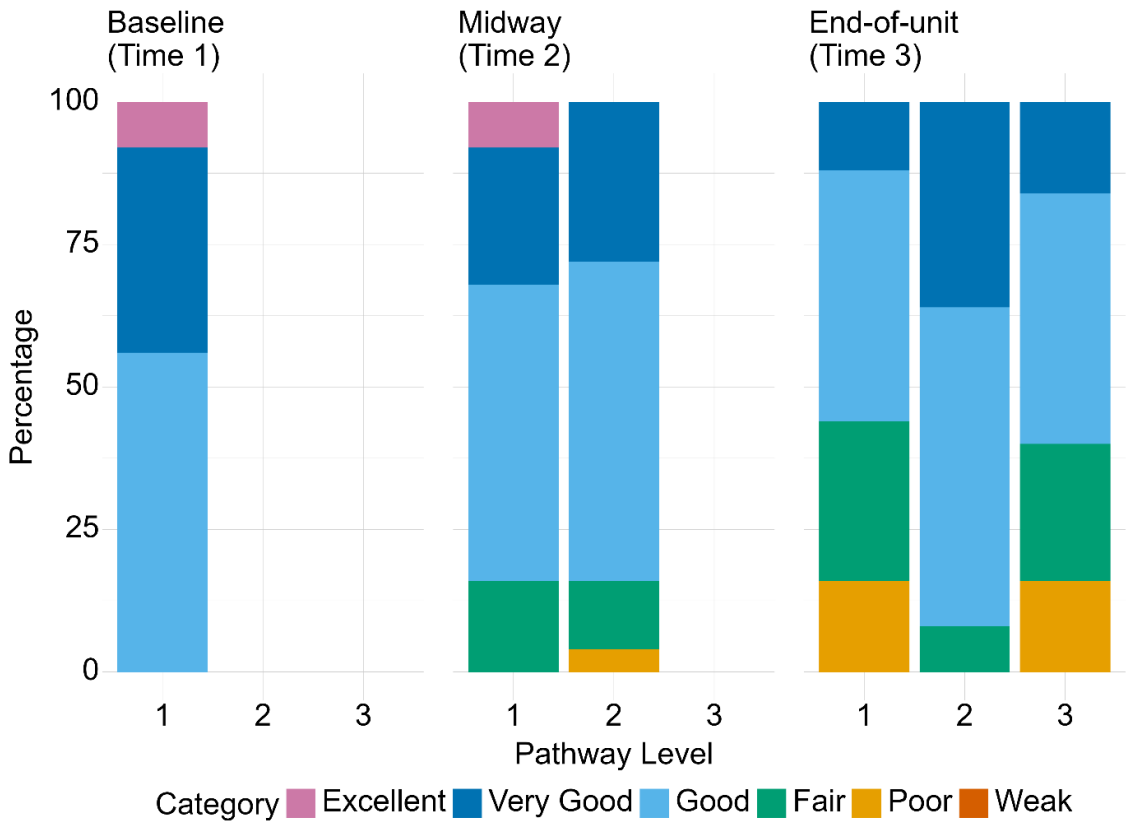
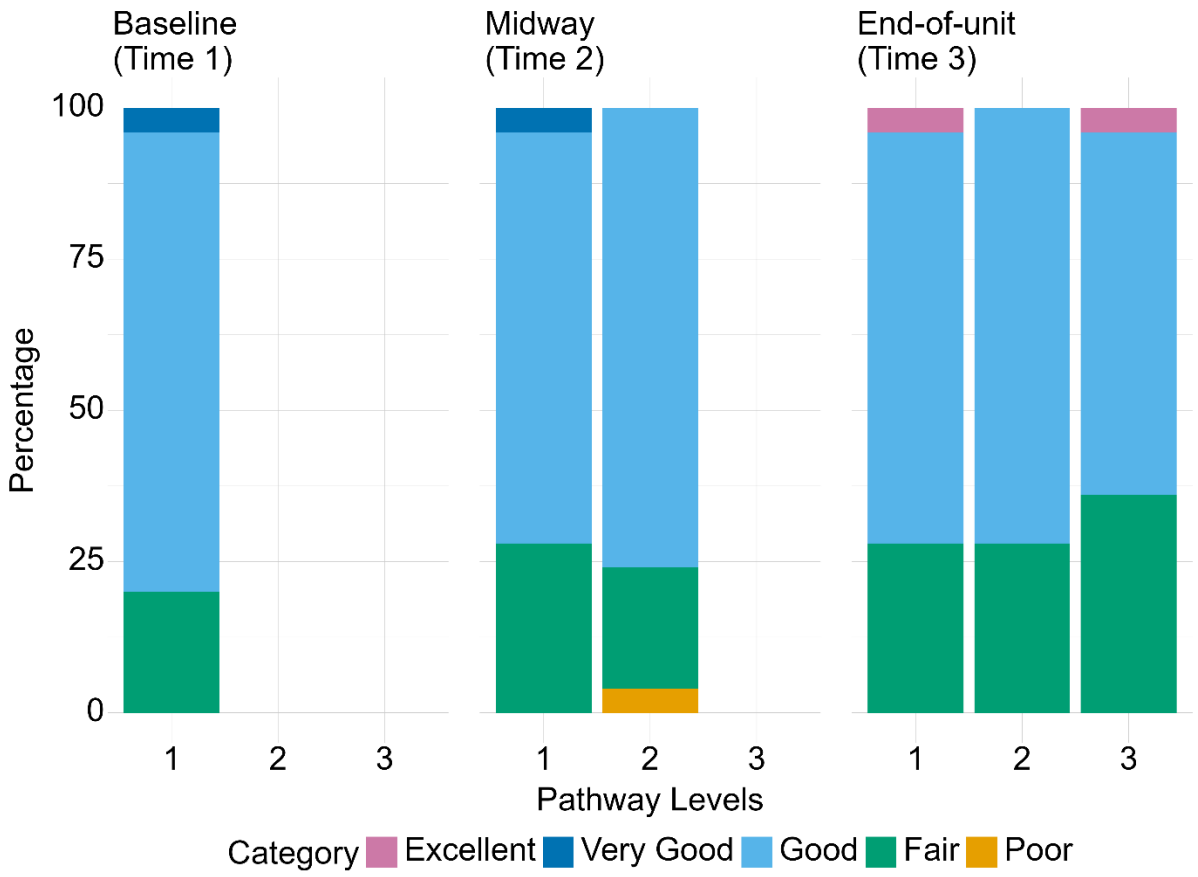


Figure 5 reports similar information for the classification consistency metric. Level 1 demonstrated *good* levels of consistency at each time point (baseline, midway, end-of-unit) for each cycle during the instructionally embedded window, with a small number of estimates demonstrating *very good* or *excellent* consistency. Level 2 showed similar patterns, with most pathway levels exhibiting *good* levels of consistency at midway and end-of-unit time points. For Level 3, most pathway levels demonstrated *good* levels of consistency. However, a smaller number of pathway levels did demonstrate *fair* or *excellent* levels of consistency, indicating slightly more variability compared to the classification accuracy findings. Overall, 149 pathway levels and time points (99%) demonstrated at least *fair* classification consistency. The findings suggest that scores produced by the TDCM-IE models were reliable overall, although there are opportunities for some improvement in the precision of Level 1 estimates at the end-of-unit assessment.

Figure 5

Estimated Classification Consistency Metrics for TDCM-IE Models



3.2.3.2 TDCM-Combined classification accuracy and consistency

Figure 6 reports the number of pathway levels within each classification accuracy category for each time point. Level 1 and Level 2 exhibited strong accuracy from baseline to end-of-unit, with estimates primarily demonstrating *good* and *very good* levels of accuracy. However, estimates trended downward into the *poor* and *fair* levels of accuracy at the final assessment. Level 3 exhibited *good* levels of accuracy at end-of-unit but *poor* and *fair* levels of accuracy at the final assessment. This result is likely because of the pilot test design. As reported previously, the final component included only two items for Level 3 of each standard, leaving relatively little information about the fourth time point on which to base classifications. The four-time-point TDCM also estimates transitions for three pathway levels at every time point, which increases the number of parameters that must be estimated, compared with simpler models. When evidence is limited, this greater model complexity can reduce the precision of parameter estimates. In addition, because the final assessment provides data only for Level 3, the model must estimate mastery for Level 1 and Level 2 at the fourth time point without direct observations for those levels. These factors—the sparse item pool for the final assessment, the need to estimate a larger

set of transition parameters, and the absence of level-specific data for Level 1 and Level 2—offer a plausible explanation for the decline in classification accuracy at the final assessment. Therefore, it is not surprising that we see poor model performance at this time point. Overall, 164 pathway level and time point combinations (73%) demonstrated at least *fair* classification accuracy.

Figure 6

Estimated Classification Accuracy Metrics for TDCM-Combined Models

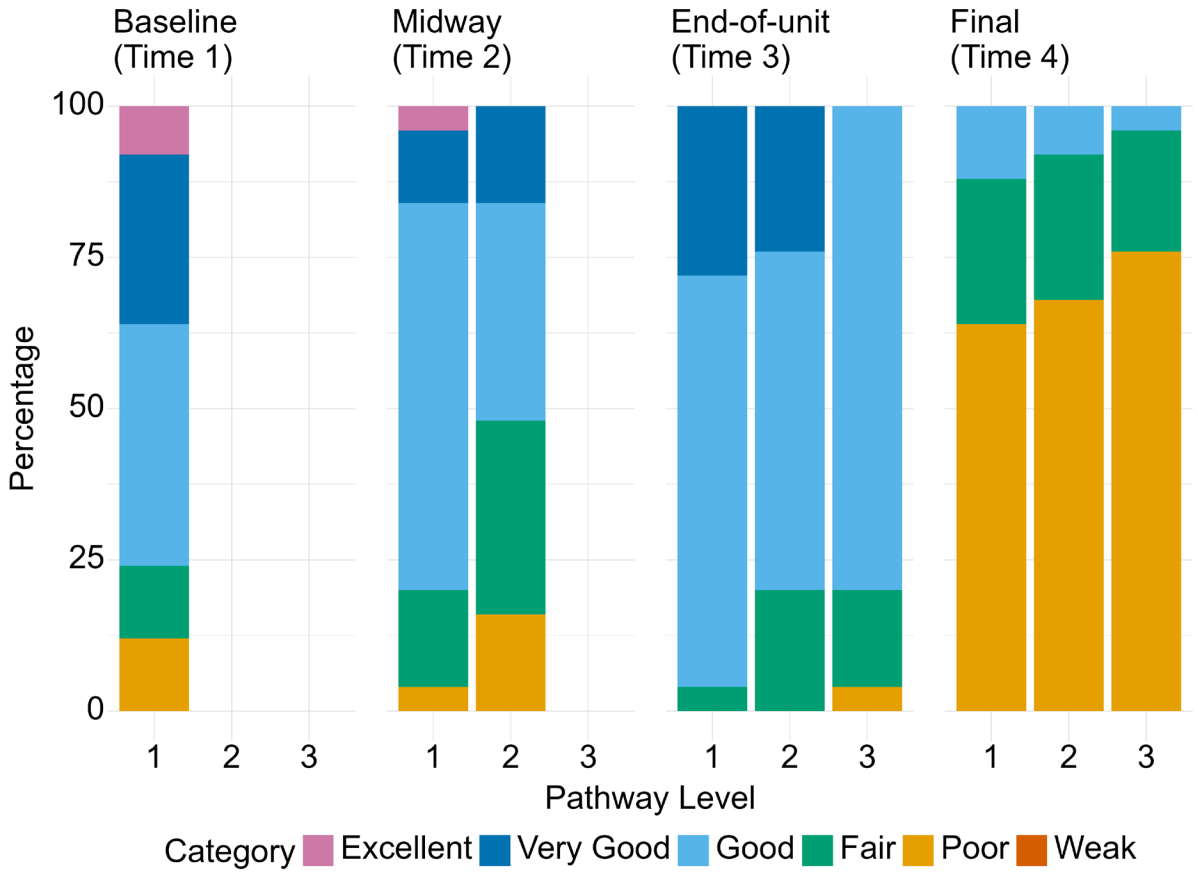
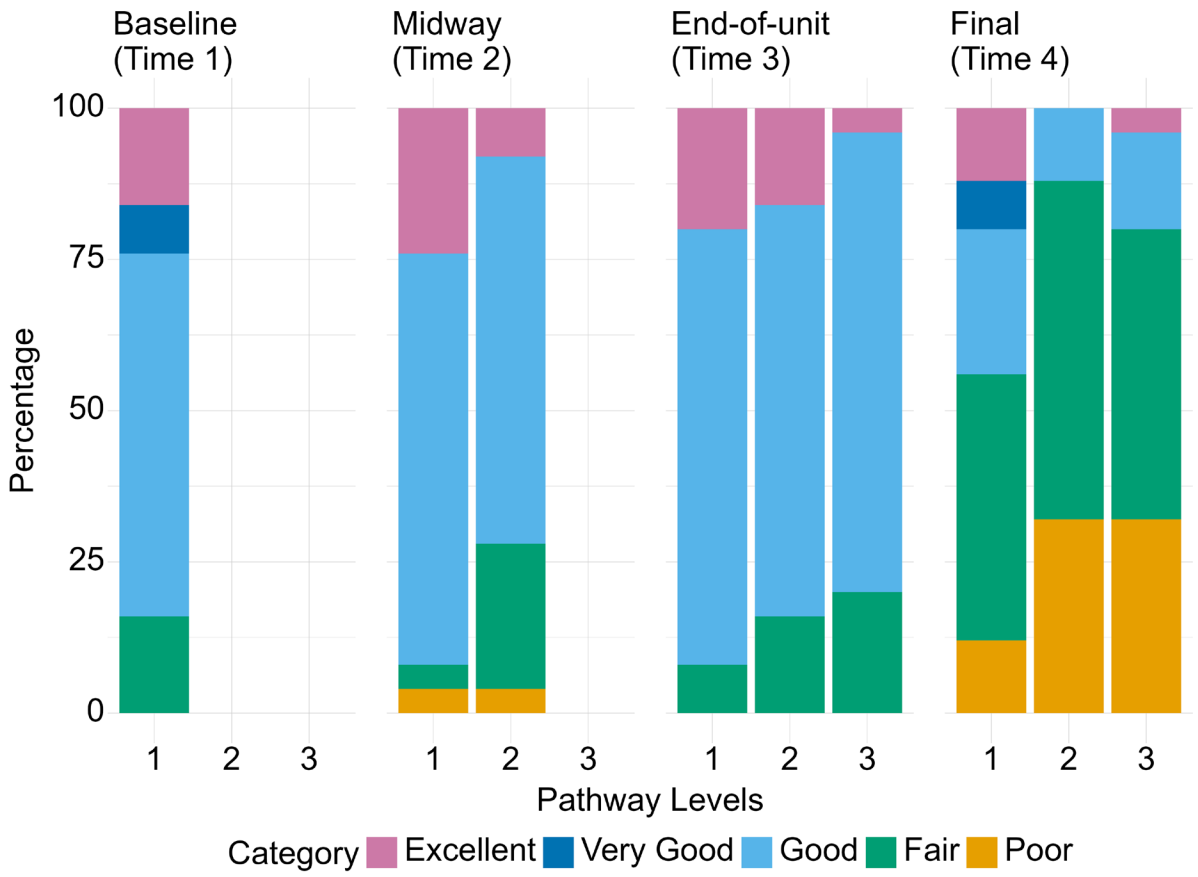


Figure 7 reports similar information for the classification consistency metric. Level 1 demonstrated the highest consistency from baseline through end-of-unit, with most estimates labeled as *good* or *excellent*. Estimates declined in terms of consistency at the final assessment, with more levels indicating *poor* and *fair* levels of consistency. Level 2 followed a similar pattern, with strong consistency at midway and end-of-unit but shifts toward lower classification consistency categories at the final assessment. Level 3 estimates were consistently *good* at end-of-unit but again reflected a decrease in consistency at the final assessment, with more pathway levels categorized as *poor* or *fair* levels of consistency. As noted earlier, the decline in model performance at the final assessment likely reflects the test design, where limited data for the final assessment

administration and the need to estimate transitions through the learning pathway increased the models' sensitivity to sparse evidence and lower classification consistency. Overall, 204 pathway level and time point combinations (91%) demonstrated at least *fair* classification consistency.

Figure 7

Estimated Classification Consistency Metrics for TDCM-Combined Models



3.2.3.3 TDCM-L3 classification accuracy and consistency

In general, strong and consistent patterns in reliability were observed across the end-of-unit and final assessment time points. Figure 8 reports the percentage of pathway levels within each classification accuracy category for each time point. Level 3 estimates demonstrated an even distribution of *fair*, *good*, and *very good* levels of accuracy at end-of-unit. At the final assessment, variability in accuracy levels declined as estimates were more heavily concentrated into *fair* and *good* levels of accuracy. While both the four-time-point, TDCM-Combined model and the two-time-point, TDCM-L3 model rely on the same limited number of items at the final assessment, the two-time-point models focused solely on Level 3 skills and estimate only a single transition. This narrower scope reduces the number of parameters to be estimated and eliminates the need to model lower-level

mastery at the final assessment time point when no direct evidence is available. These features reasonably explain why classification accuracy remained stable for the TDCM-L3 models despite the limited final assessment data. Overall, 47 pathway levels (94%) demonstrated at least *fair* classification accuracy over all testing occasions and models.

Figure 8

Estimated Classification Accuracy Metrics for TDCM-L3 Models

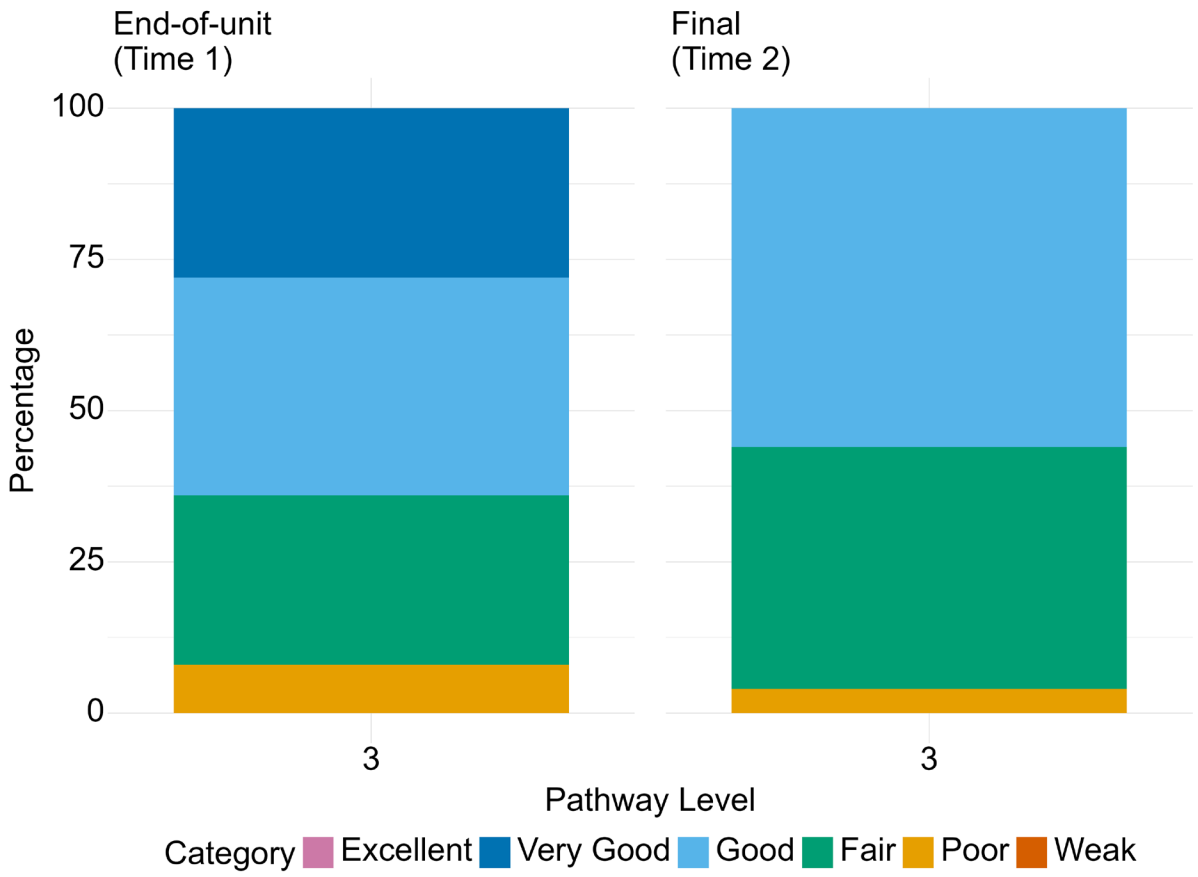
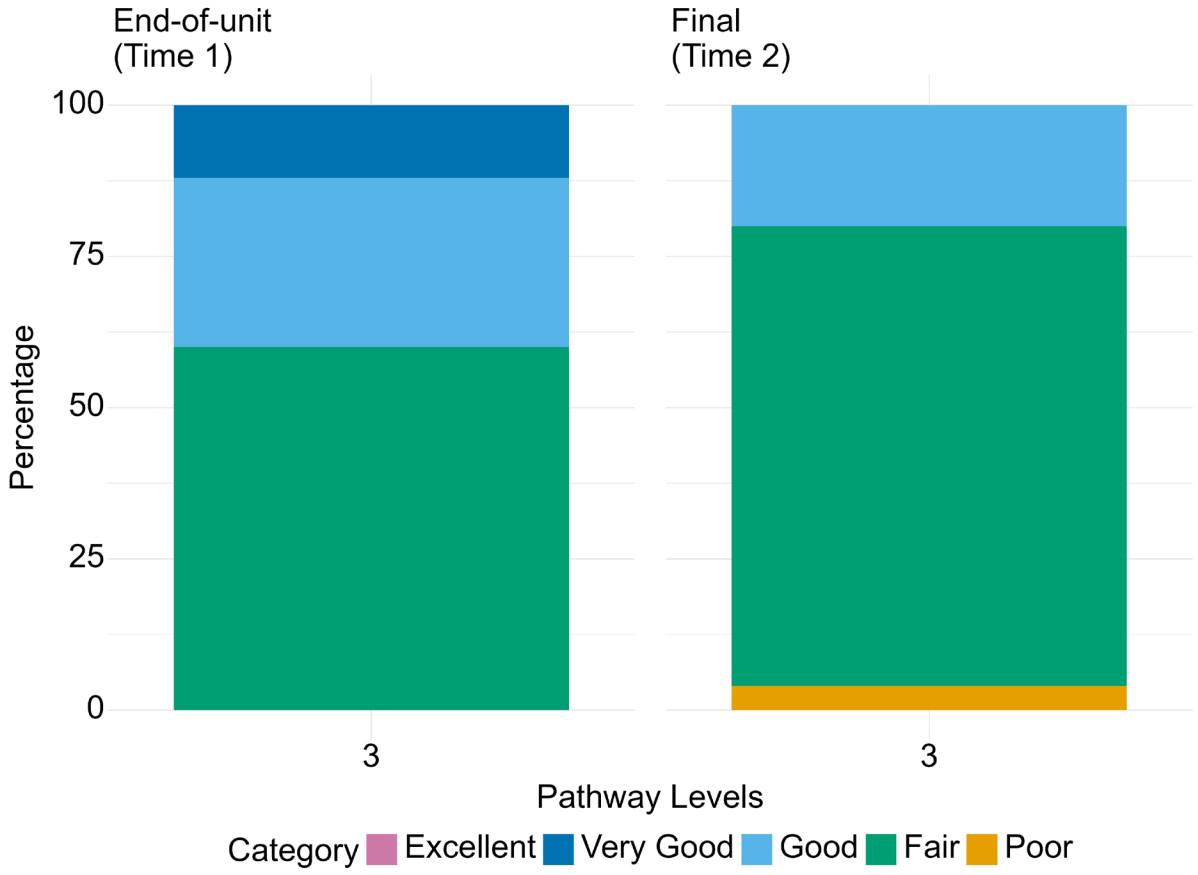


Figure 9 reports similar information for the classification consistency metric. Level 3 estimates demonstrated mostly *fair* (15) and *good* (7) levels of consistency at end-of-unit, with a few estimates designated as *very good*. At the final assessment, estimates were more concentrated into *fair* categories, with a few estimates demonstrating *good* levels of consistency. Overall, 49 pathway levels (98%) demonstrated at least *fair* classification consistency over all testing occasions and models. These results align with the accuracy findings, indicating that classification consistency remained high across time points and reinforcing the overall reliability of Level 3 mastery classifications despite the limited final assessment pool.

Figure 9

Estimated Classification Consistency Metrics for TDCM-L3 Models



3.2.4 Summary of DCMs in Through-Year Designs

Findings from the 2024–2025 PIE pilot study indicate that transition DCMs that incorporate only instructionally embedded data provided sufficient information to produce fine-grained mastery scores that reflect student learning by the end of the year. Adding final assessment data contributed little additional information and was associated with lower precision for some mastery estimates. This outcome likely reflects design features of the

current system, notably the limited number of items administered per standard on the final assessment.

3.3 Hybrid DCM/IRT Models

The final class of models we consider in this report are hybrid models, which are designed to balance the strengths of DCM and IRT approaches. These models incorporate fine-grained mastery scores from DCMs with scale-based reporting in an IRT framework. Together, these features make hybrid models a promising option for through-year assessment systems that aim to serve both accountability and instructional purposes. Specifically, we explore the use of Beta-IRT models (Noel & Dauvier, 2007; Zopluoglu & Lockwood, 2024). These models are similar to traditional IRT models but take values bounded by 0 and 1 as input (e.g., probabilities), rather than categorical data (e.g., dichotomous or polytomous responses), as in the discussion of 2PL and GRM models in the Item Response Theory section. The Beta-IRT model is ideally suited to a hybrid DCM/IRT approach, as the outputs of the DCM component are probabilities that students have mastered each attribute. Thus, the mastery profile can be built using a DCM, as described in the Diagnostic Classification Models sections, and then the scale score can be directly inferred from the profile using the Beta-IRT model.

3.3.1 Data Conditions

To evaluate how different sources of evidence influence the performance of the hybrid models, we estimated four variants that reflect two key design decisions. The first decision concerns final assessment information: whether mastery probabilities from the instructionally embedded assessment are used alone or combined with item responses from the final assessment component. The second decision concerns attribute scope: whether mastery probabilities for all three pathway levels are included or the precursor Level 1 skills are excluded (i.e., only Levels 2 and 3 contribute to the scale score).

These two factors yield four models. Hybrid-IE uses mastery probabilities from all instructionally embedded assessment—baseline, midway, and end-of-unit—within a Beta-IRT framework. Hybrid-IE-noL1 is identical but excludes Level 1 mastery probabilities. Hybrid-Combined incorporates both instructionally embedded mastery probabilities and final assessment item responses in a joint Beta-IRT/GRM framework. Hybrid-Combined-noL1 applies the same approach but omits Level 1 mastery probabilities.

Analyses of the TDCM models indicate that inclusion of final assessment data within the longitudinal modeling framework resulted in inadequate technical properties for mastery scores estimated at the final assessment. Therefore, hybrid approaches using final assessment mastery scores were not pursued. This outcome should not be interpreted as evidence that the final assessment data are uninformative; rather, it reflects challenges in fitting longitudinal DCMs that maintain reliability for the final assessment classifications. The final assessment remains a viable option for inclusion within a summative modeling framework (i.e., where final assessment items directly inform the continuous latent trait in the Beta-IRT model; Table 7).

Table 7

Hybrid DCM/IRT Models and Data Conditions

Model	DCM mastery probability data	IRT scale score data	Levels included
Hybrid-IE	Baseline, midway, end-of-unit	DCM probabilities only	Level 1, Level 2, Level 3
Hybrid-IE-noL1	Midway, end-of-unit	DCM probabilities only	Level 2, Level 3
Hybrid-Combined	Baseline, midway, end-of-unit	DCM probabilities + final items	Level 1, Level 2, Level 3
Hybrid-Combined-noL1	Midway, end-of-unit	DCM probabilities + final items	Level 2, Level 3

3.3.2 Model Definitions

The Beta-IRT model (Noel & Dauvier, 2007) extends traditional IRT to support continuous-bounded responses on the 0–1 interval such as, proportions, probabilities, or other continuous outcomes, rather than strictly dichotomous- or polytomous-scored data. Unlike the 2-parameter logistic (2PL) or graded response model (GRM), which are restricted to discrete data, the Beta-IRT framework assumes that observed responses follow a beta distribution. By modeling responses with a beta distribution, the framework can describe skewed patterns of responses and capture the variability in responses around the conditional mean (Zopluoglu & Lockwood, 2024).

In the Beta-IRT model, a discrimination parameter controls the rate at which the conditional mean of the beta distribution increases with ability, paralleling the slope parameter in the 2PL IRT model. A difficulty parameter shifts the location of the response curve along the ability scale. Finally, a dispersion parameter controls the spread of responses around the conditional mean, allowing the model to represent heterogeneity in response precision across items. The conditional mean of the Beta-IRT model has the same functional form as the logistic IRT model, while the conditional variance is directly influenced by both the conditional mean and the dispersion parameter. Together, these properties make the Beta-IRT model a flexible framework for assessment models that require both mastery-based interpretations and summative scores. By capturing continuous, probabilistic evidence of learning, the model supports the aggregation of diverse assessment outcomes into interpretable and reliable measures for both instructional and accountability purposes.

In addition to a standard Beta-IRT model, we can also define a model that uses both probabilities and traditional item responses. This model combines features of the traditional 2PL and GRM frameworks and the Beta-IRT framework to accommodate mixed response formats within a single latent trait model. In this specification, all items load onto a single, continuous latent trait, with dichotomous items contributing information through a 2PL likelihood and continuous-bounded items contributing information through the Beta-

IRT likelihood. This design enables the model to leverage both traditional item responses and probabilistic evidence.

3.3.3 Model Estimation

The hybrid models were estimated using the same Bayesian framework as the IRT models, employing Markov chain Monte Carlo procedures in *Stan* (Carpenter et al., 2017) using *rstan* (Stan Development Team, 2025). This approach provides flexibility in accommodating the mixture of discrete and continuous-bounded likelihoods while supporting robust evaluation of model fit (Thompson, 2019).

For the 2PL/GRM component, item intercept parameters were assigned normal prior with mean 0 and variance 2. Item slope parameters were assigned lognormal priors with mean 0 and variance 1, and the item slope parameters were constrained to be positive to preserve monotonicity. For the Beta-IRT component, the item intercept parameters were assigned normal priors with mean 0 and variance 2, and the slope parameters were given lognormal priors with mean 0 and variance 1, also constrained to be positive. The item dispersion parameters were assigned normal priors with mean 0 and variance 2. Across both components, ability parameters were jointly estimated with item parameters and were anchored by a standard normal prior, ensuring that student scale scores were estimated on a common metric.

Convergence diagnostics followed the same criteria as for the IRT models (Vehtari et al., 2021), requiring all $\hat{R}$ values to be less than 1.01 and effective sample sizes to be greater than 400. After confirming convergence across the four sampling chains, posterior means were used to summarize parameter estimates.

3.3.4 Model Evaluation

Posterior predictive model checking (PPMC) was used to assess how well the hybrid models reproduced observed response patterns. For models based solely on discrete item responses, fit was evaluated by comparing the observed raw-score distribution with distributions generated from replicated datasets. For hybrid models that incorporated mastery probabilities, this procedure was adapted by aggregating each student's probabilistic responses into a pseudo-raw score (i.e., the sum of their probabilistic scores across items) and rounding that sum to the nearest integer to create a distribution comparable to the discrete response models. Although this approach provides a practical extension of PPMC to hybrid frameworks, discretizing continuous data inevitably entails some loss of information and should be viewed as an approximation of absolute fit.

Posterior predictive *p*-values (*ppp*) were calculated to summarize fit across the two study factors: inclusion of final assessment data and inclusion of Level 1 evidence. Models that excluded the final assessment component closely reproduced the observed data but produced *ppp* values near 1.0, suggesting potential overfitting and limited generalizability to other testing designs. In contrast, models that incorporated final assessment data yielded *ppp* values near 0, indicating lack of fit. Little difference was observed between

models that included or excluded Level 1 evidence, pointing to the presence of final assessment data as the primary factor affecting model fit in the current study.

Several features of the data may help explain these results. The traditional IRT model performed poorly when fit to the final assessment responses. Adding 50 final assessment items to the 75 mastery probabilities in the Hybrid-Combined model effectively introduced a large set of data that was difficult to model with the 2PL and GRM frameworks. In addition, the mastery probabilities included in both the Hybrid-IE and Hybrid-Combined models were heavily concentrated near 0 or 1, and these boundary cases can influence estimation in models that are based on a Beta distribution. When values are clustered near the boundaries, the associated shape parameters move toward extreme values, creating sharply peaked likelihoods and increasing the sensitivity of parameter estimates. Together, the poor fit of the final assessment responses using a standard IRT framework and the presence of boundary cases in the mastery probability data provide a plausible explanation for why models that incorporated final assessment data demonstrated weaker absolute fit, whereas models limited to instructionally embedded evidence exhibited potential overfit (Table 8).

Table 8

Item Response Theory Model Fit Evaluation

Model	Posterior predictive *p*-value	Fit evaluation
Hybrid-IE	1.000	Overfit
Hybrid-IE-noL1	1.000	Overfit
Hybrid-Combined	.000	Poor
Hybrid-Combined-noL1	.000	Poor

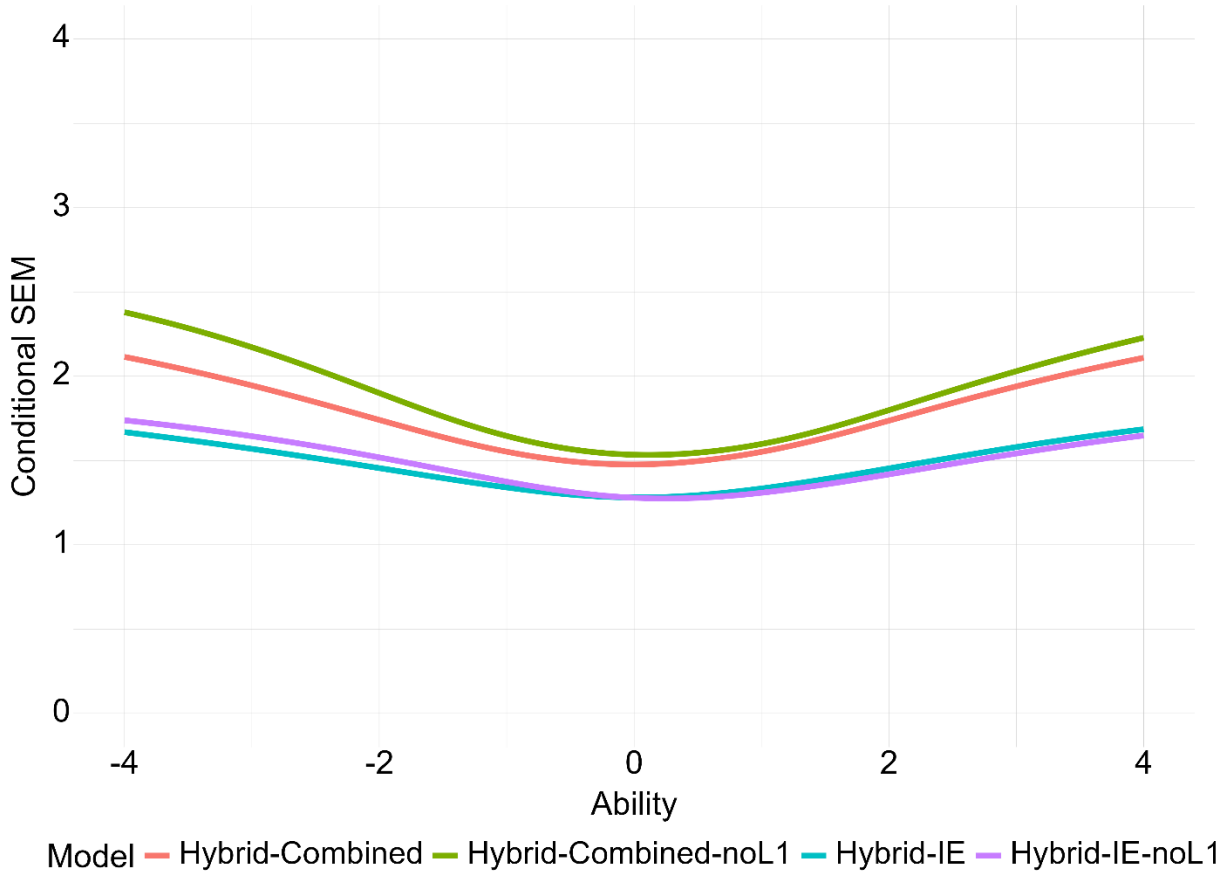
As with the traditional IRT models, we also can estimate the precision of ability estimates in the Beta-IRT model through the test information function (TIF). For the Beta-IRT model, the TIF is generalized to continuous-bounded data, capturing how measurement precision varies across the ability scale while accounting for the shape of the Beta distribution underlying each mastery result. For technical details on the derivation of the TIF, see Noel and Dauvier (2007).

Clear differences in reliability were observed between models that included final assessment data and models that relied solely on instructionally embedded evidence. Figure 10 shows reliability, expressed as conditional standard errors of measurement (CSEM) across the latent trait continuum, for each hybrid model. After standardizing by test length, both Hybrid-IE variants, whether using mastery probabilities for Levels 1 – 3 (Hybrid-IE) or restricted to Levels 2 – 3 (Hybrid-IE-noL1), consistently provided lower CSEM than their Hybrid-Combined counterparts. Although adding final assessment items increased the total item pool, the Hybrid-Combined models did not show improved reliability after results were scaled for test length. Instead, the inclusion of the final assessment component appeared to reduce precision in this case study, indicating that

the instructionally embedded assessments alone supplied sufficient information to capture the targeted constructs. Differences between models that included versus excluded Level 1 mastery probabilities were minimal, suggesting that the decision to omit precursor Level 1 evidence had little effect on the reliability of the resulting summative scale scores.

Figure 10

Conditional SEM of Hybrid DCM/IRT Models, Scaled by Test Length



3.3.5 Summary Hybrid DCM/IRT Models in Through-Year Designs

Results from the 2024–2025 PIE pilot administration show that hybrid DCM/IRT models that incorporate mastery evidence solely from the instructionally embedded window provided a stronger fit and reliability than models that combined instructionally embedded evidence with final assessment responses. Hybrid models that incorporated final assessment data exhibited weaker performance under these conditions and therefore warrant additional study before their use in similar settings can be fully supported.

3.4 Summary and Considerations of Summative Models

Results of the modeling analyses using data collected from the PIE pilot study show that at least one variant of each modeling framework (IRT, DCM, hybrid) met minimum standards for technical adequacy, with the strongest performance observed for models that drew evidence primarily from the instructionally embedded assessments. These results illustrate the potential of multiple psychometric approaches to combine information from repeated testing occasions into a single indicator of student performance by the end of the year.

Equally important, the analyses highlight how strongly model performance depends on assessment design. Key factors such as the number of items and the breadth of content coverage influence the reliability and interpretability of resulting scores. For example, increasing the size of the final assessment item pool may improve the precision of year-end estimates; however, doing so may undermine the goals and intended outcomes of a through-year model. Additionally, testing programs should consider the levels on which students are assessed. In the PIE pilot, students were required to test on all pathway levels within each standard. Future implementations may choose a more adaptive design in which students begin at a Level 2 or Level 3 skill and performance on previous standards informs placement on future standards. However, this type of design creates a sparse matrix with intentionally missing data (i.e., no Level 1 data for some students), which may complicate the estimation of a longitudinal model. Therefore, it is important to consider what type of reporting is desired and what level of granularity for precursor skills is useful for educators and students to make instructional decisions. Finally, extending content beyond grade-level expectations may place student performance along a broader learning continuum (e.g., precursor or successor skills that may be more closely aligned to content standards in adjacent grades).

Overall, these findings underscore that the success of through-year IRT, DCM, and Hybrid models relies less on the choice of framework than on the structural and content decisions made during test design. These observations are specific to the present design in which the final assessment component includes a limited number of items per standard. In other through-year settings, particularly where the final assessment includes a broader set of items or follows a different content design, incorporating final assessment data into a hybrid framework may yield different findings and potential benefits. The current findings are intended to illustrate how model performance interacts with design features, underscoring the importance of aligning the choice of hybrid model with the goals and structure of the assessment system. Future implementations should continue to research how different designs, particularly those that involve a final assessment component, can best support the goals and claims of through-year assessment programs.

Finally, we note that choosing a through-year test design should be determined by the claims that an assessment system is intended to support. The choice of a scoring model, then, should be based on both these claims and the chosen test design. However, other

factors also need to be considered. The next section describes additional policy concerns that influence the choice of both test design and scoring model.

4 Policy Considerations for Summative Scoring Models

As described in the previous sections, there are many psychometric choices and test design considerations when developing a summative model for a through-year assessment. However, a summative scoring model does not exist in isolation. Rather, the results are required to feed into state accountability systems. This includes the division of student achievement into performance levels that are used for an academic achievement indicator and another indicator, which is usually a measure of student growth (U.S. Department of Education, 2017). Given the differences in the types of results that IRT, DCM, and hybrid models support, these summative models similarly support different types of standard setting methods for determining performance levels and growth models that support growth measures. Thus, the choice of a summative model should also consider how these elements of an accountability system will be addressed.

4.1 Standard Setting Methods for Summative Models

If an IRT model is used for the scoring model, well-known and widely used standard setting methods are available. For example, the assessment system may use a bookmark (Lewis et al., 2012) or Angoff (Plake & Cizek, 2012) method, both of which are used to set cut points in the overall scale score. Similarly, a longitudinal DCM has existing methods to set achievement standards from a profile of mastered skills. For example, one may use the condensed mastery profiles method to set a cut point in the total number of mastered skills (Clark et al., 2017).

The hybrid IRT/DCM summative models may implement standard setting methods consistent with the IRT or DCM approach; however, modifications are required, and additional research is needed to evaluate the properties of performance levels derived from the methods. For example, the bookmark method normally uses an ordered item booklet. Because the "items" in the hybrid approach are the DCM mastery probabilities, a bookmark method for a hybrid model may use an ordered "skill booklet." With a skill booklet, panelists determine the point at which a minimally qualified candidate demonstrates mastery, rather than provide a correct response. Similarly, a profile-based standard setting method, like condensed mastery profiles, may be used. Because the hybrid models also produce scale scores, reports may include scale score ranges associated with each performance level, which the profile itself determines.

4.2 Growth Models for Summative Models

As with the standard setting methods, there are well-known growth models that are compatible with an IRT scoring model such as categorical gains, student growth percentiles, and value-added models, among others. For DCM scoring models, growth models using performance level distributions (e.g., categorical gains, transition tables) also apply. However, methods that use a continuous measure of growth would require

further research. For example, an assessment system needs to connect attributes (or learning pathways) across grades to vertically articulate the progression of skills across grades (e.g., Level 3 skill in grade 5 may be a Level 2 skill in grade 6). Additional work is also needed to translate the outputs of the growth model into a growth measure. In a DCM-based assessment, growth is indicated by the mastery of additional skills. However, because these are categorical mastery classifications, it is not clear how much "growth" a change in mastered skills represents. For example, if two skills are very close to each other conceptually, relatively little additional achievement may be needed to demonstrate mastery of the additional skill. On the other hand, if two skills are conceptually distinct, great deal of additional achievement may be needed to demonstrate mastery of a second skill. In both cases, one additional skill is mastered, but different amounts of growth are required.

Because the hybrid IRT/DCM models produce a scale score, existing growth models may be applied to the hybrid models, as they are for IRT models (e.g., categorical gains, student growth percentiles, value-added models). The scale score also simplifies the creation of a growth measure from the outputs of the growth models, as the scale score itself defines the amount of growth. However, like the DCM models, further work would be needed to define methods for creating a vertical scale from the measured skills and evaluate the properties of the scale to ensure the psychometric properties are technically adequate.

In summary, there are options for reporting growth in all summative models investigated in the PIE project. However, if an organization is committed to a particular growth model, that position may limit the available scoring models.

5 Conclusions

In summary, the development of a summative scoring model for a through-year assessment is a multifaceted process that involves a combination of psychometric, test design, and policy considerations. Because of the interrelated nature of these considerations, it is critical that assessment systems clearly define the claims that must be supported for the intended uses. For the PIE system, these claims are defined in the theory of action. Returning to Figure 2, the choice of a summative model affects several claims related to student results. Table 9 summarizes these claims.

Table 9

Support for Relevant Claims in the PIE Theory of Action Provided by Each Summative Model

Claim	IRT	DCM	Hybrid
I: Mastery results represent what students know and can do relative to the learning pathways.	**Not supported**	**Results reported directly as the set of mastered knowledge, skills, and understandings (KSUs)**	**Mastery results directly inform summative scale score**
K: Summative results accurately reflect student achievement of grade-level academic content standards.	**Supported with a single scale score**	**Supported with a profile of mastered KSUs**	**Supported with both scale score and diagnostic profile**
L: Educators make instructional decisions based on data from the PIE assessments.	**Not well suited to instructional decision-making**	**Instructional decision-making based on mastery profile**	**Instructional decision-making based on mastery profile**
M: Students make progress toward mastery of grade-level content standards.	**Supported with existing growth models**	**Additional research needed to evaluate profile-based growth**	**Supported with existing growth models**

Note. IRT = item response theory; DCM = diagnostic classification model; KSUs = knowledge, skills, and understandings.

For PIE, Claim I and Claim L, specifically, are related to fine-grained mastery results that are relevant for instructional decision-making. IRT models are not well suited to providing these types of results, whereas DCMs and hybrid models, which use DCMs to provide results during instructionally embedded testing, are ideal for this kind of reporting (Thompson & Clark, 2024). Additionally, DCMs currently have limitations for estimating student growth (Claim M) with noncategorical models (e.g., student growth percentile, value-added models), which would require additional research.

Using an evaluation of the claims in the PIE theory of action, a hybrid IRT/DCM model would be the most appropriate for the current iteration of the system's design. However, it is important to recognize that different test designs or claims could reasonably result in a different choice of scoring model. The choice of a summative model should be driven not just by psychometric adequacy, but also by consistency with the theory of action. In this report, we described three classes of psychometric models and explored how to use those models to support summative scoring in a through-year assessment system. In the end, the success of any through-year summative model depends on its coherence with the system's overarching goals, technical defensibility, and usefulness for decision-making. Maintaining alignment among these elements will ensure that the assessment system supports valid interpretations of student learning across the year.

References

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–424). Addison-Wesley.

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

Clark, A. K., Nash, B., Karvonen, M., & Kingston, N. (2017). Condensed mastery profile method for setting standards for diagnostic assessment systems. *Educational Measurement: Issues and Practice, 36*(4), 5–15. https://doi.org/10.1111/emip.12162

Dadey, N., Evans, C. M., & Lorié, W. (2023). *Through-Year Assessment: Ten Key Considerations*. National Center for the Improvement of Educational Assessment. https://files.eric.ed.gov/fulltext/ED627081.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/

Gong, B. (2025, June 23–25). Through-year assessments: Hopes and dilemmas. In *Emerging lessons from multiple states on validity and technical issues of through-year assessments: Guidelines for assessment programs* [Symposium]. National Conference on Student Assessment, Denver, CO.

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

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

Kim, E. M., Nash, B., & Swinburne Romine, R. (2024). *Pathways for instructionally embedded assessment (PIE): Developing learning pathways for the PIE assessment system*. https://pie.atlas4learning.org/sites/default/files/documents/resources/Developing_Learning_Pathways_for_the_PIE_Assessment_System.pdf

Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, E. M. (2012). The bookmark standard setting procedure. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 225–253). Routledge.

Madison, M. J., & Bradshaw, L. P. (2018). Assessing growth in a diagnostic classification model framework. *Psychometrika, 83*(4), 963–990. https://doi.org/10.1007/s11336-018-9638-5

Marion, S., D'Brot, J., & Brandt, C. (2022). *Assessment design and implementation considerations for the Montana Alternate Student Testing (MAST) pilot program*. Montana Office of Public Instruction. https://www.nciea.org/wp-content/uploads/2022/08/MAST-Report-R4.pdf

Missouri Department of Elementary and Secondary Education. (2025). *Innovation assessment demonstration authority for Missouri Success Ready Student Assessment, condensed version*. https://dese.mo.gov/quality-schools/assessment/media/pdf/final-iada-condensed-version-feedback

Munyan-Penney, N., & Barone, C. (2020). *Innovative assessment pilot: Louisiana's LEAP 2025 humanities assessment*. Education Reform Now. https://edreformnow.org/wp-content/uploads/2020/11/Louisiana-IADA-LEAP-2025-Humanities.pdf

Noel, Y., & Dauvier, B. (2007). A beta item response model for continuous bounded responses. *Applied Psychological Measurement*, *31*(1), 47-73.

Plake, B. S., & Cizek, G. J. (2012). Variations on a theme: The modified Angoff, extended Angoff, and yes/no standard setting methods. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 181–199). Routledge.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. The Guilford Press.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 34*(Suppl 1), 1–97. https://doi.org/10.1007/BF03372160

Stan Development Team. (2025). *RStan: The R interface to Stan*. (R package version 2.32.7) [Computer software]. 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., & Clark, A. K. (2024). Improving instructional decision-making using diagnostic classification models. *Educational Measurement: Issues and Practice*, *43*(4), 146–156. https://doi.org/10.1111/emip.12619

Thurgood, L., Walter, E., Carter, G., Henn, S., Huang, G., Nooter, D., Smith, W., Cash, R. W., & Salvucci, S. (2003). *NCES handbook of survey methods* (Technical report). U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. https://nces.ed.gov/pubs2003/2003603.pdf

U.S. Department of Education. (2017). *Every Student Succeeds Act—Accountability, state plans, and data reporting: Summary of final regulations*. https://www.ed.gov/sites/ed/files/policy/elsec/leg/essa/essafactsheet170103.pdf

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

Zopluoglu, C., & Lockwood, J. R. (2024). A comparative study of item response theory models for mixed discrete-continuous responses. *Journal of Intelligence*, *12*(3), 26. https://doi.org/10.3390/jintelligence12030026