

Beyond “Three-in-a-Row”: Toward Fairer Adaptive Mastery Decisions in AI-Supported Formative Assessment

Chunling Niu*

The University of Texas at
San Antonio
chunling.niu@utsa.edu

Kelly D. Bradley

Marshall University
kelly.bradley@marshall.edu

Brian F. Waltman

Fairmont State University
Brian.Waltman@fairmont
state.edu

Biao Ma

The University of the
Incarnate Word
martin.ma.ciw@gmail.com

Abstract

We compare the three-correct-in-a-row mastery rule with an adaptive, measurement-informed classifier using ASSISTments Skill Builder data. The adaptive model agreed with terminal outcomes far more often at early checkpoints (.787 vs. .435), but the rule remained competitive where it already performed well, supporting context-sensitive rather than wholesale replacement.

1 Introduction

Artificial intelligence is reshaping educational assessment, feedback, and pedagogy. As formative systems become increasingly adaptive, they create new demands for evidence about what these systems measure, how consistently they classify learners, and how their decisions affect instruction. Within this broader shift, mastery classification deserves renewed attention. In many digital learning systems, mastery decisions determine whether students advance, continue practicing, or receive additional support. These are consequential measurement decisions, not merely software settings.

One longstanding approach to mastery classification in intelligent tutoring systems is the use of simple deterministic rules. A prominent example is the three-correct-in-a-row criterion, under which a learner is considered to have mastered a skill after answering three consecutive items correctly (Heffernan & Heffernan, 2014; Pardos et al., 2013). This rule is attractive because it is easy to implement, transparent, and simple to communicate to educators and developers. However, its simplicity may also be its limitation. A fixed streak criterion ignores other potentially important indicators of learner status, such as repeated attempts, hint reliance, response timing, and contextual differences across skills and item templates, that probabilistic learner models were designed to exploit (Corbett & Anderson, 1995).

This paper investigates whether a more adaptive, measurement-informed mastery framework can improve upon the traditional three-correct-in-a-row rule. The aim is not simply to show that a predictive model can outperform a heuristic, but to address a measurement-science question: does a richer, adaptive mastery criterion yield more consistent and instructionally defensible decisions than a fixed rule? Using the publicly available ASSISTments Skill Builder corrected dataset, the study compares the two criteria with a focus on validity, reliability, and instructional effectiveness in AI-supported formative systems.

2 Background and Problem Statement

Traditional intelligent tutoring and formative assessment systems have often relied on operationally convenient mastery rules (Corbett & Anderson, 1995; Heffernan & Heffernan, 2014). The three-correct-in-a-row criterion classifies mastery using only consecutive correctness. Although useful in structured settings, such a rule may produce unstable or context-sensitive decisions because it does not account for additional learner evidence that is routinely available in digital environments.

The consequences of such instability matter. If mastery is declared too early, students may advance without adequate understanding; if too late, they may be held in unproductive practice, or “wheel-spinning” (Beck & Gong, 2013). In either case the issue is not merely predictive accuracy, but the measurement quality and instructional defensibility of the mastery decision itself.

This is especially relevant as formative systems incorporate AI-supported decision-making. Adaptive systems can draw on multiple indicators of learner interaction and may offer a stronger evidentiary basis than a fixed rule alone, but any

such approach must itself be evaluated against rigorous measurement standards. The present study therefore treats both criteria as competing classification frameworks.

3 Purpose and Research Questions

This study evaluates the consistency and contextual robustness of the traditional three-correct-in-a-row mastery criterion and compares it with an adaptive mastery model derived from learner interaction data in ASSISTments. Four research questions guide the work:

RQ1. *How consistent are mastery decisions produced by the traditional three-correct-in-a-row rule across repeated opportunities within the same student-skill trajectory?*

RQ2. *To what extent do mastery decisions under the traditional rule vary across skills, problem templates, and instructional sequences?*

RQ3. *Does an AI-supported adaptive mastery model yield more consistent mastery classifications than the traditional rule across varying learning contexts?*

RQ4. *Does the adaptive model reduce context-related instability in mastery decisions relative to the traditional three-correct-in-a-row criterion?*

4 Data Source

The study uses the ASSISTments “Skill Builder” 2009–2010 corrected dataset, a large publicly available corpus from an online formative tutoring system (Feng, Heffernan, & Koedinger, 2009; Heffernan & Heffernan, 2014). The data contain item-level student interaction records including correctness, attempt count, first-response time, hint use, skill identifiers, problem-template identifiers, and sequence and position information. The skill-builder problem sets are themselves administered under a mastery-learning rule requiring a set number of correct responses in a row (three by default), so the traditional criterion under study is the dataset’s native mastery mechanism. Because the data capture repeated student-skill opportunities, they support both rule-based and adaptive mastery classifications.

The dataset comes from a traditional intelligent tutoring system, not a generative AI tutoring environment. The contribution is therefore not a comparison between rule-based and GenAI tutoring, but an evaluation of whether an adaptive,

AI-supported mastery framework provides a stronger decision model than a fixed rule.

5 Method

5.1 Analytic Design

Two approaches to mastery classification were compared:

1. **Traditional rule-based mastery criterion:** a student is classified as having mastered a skill after achieving three consecutive correct responses.
2. **Adaptive mastery model:** a model-based classification framework, specified in Section 5.2, that incorporates learner interaction features alongside skill, template, and sequence context.

The analytic design focused on decision consistency, contextual variation, and error reduction, rather than causal effects.

5.2 Variables and Adaptive Model Specification

The adaptive model is a supervised classifier that estimates, at each opportunity within a student-skill sequence, the probability that the sequence will resolve in mastery. It draws only on information available at or before that opportunity, so that it operates under the same informational constraints that a deployed system would face at the moment of decision.

The analyses used the following classes of variables:

- *Performance indicators:* correctness, attempt count, cumulative proportion correct, current consecutive-correct streak
- *Support indicators:* hint count, cumulative hint count, hint dependence
- *Timing indicator:* response time
- *Contextual indicators:* skill, skill name, template, sequence position
- *Comparative outcomes:* agreement with terminal outcome, checkpoint-level error rates, and contextual improvement indices

Two properties of this specification matter for interpreting the results. First, the consecutive-correct streak variable reproduces the traditional rule exactly, so the traditional criterion is nested within the evidence available to the adaptive model, and any difference in performance reflects information the rule discards rather than information it cannot access in principle. Second,

cross-validation was grouped by student, so all sequences belonging to a learner fell within a single fold and all reported metrics were computed on held-out cases. Without student-level grouping, repeated opportunities from the same learner would appear in both training and evaluation sets and inflate apparent performance.

The classifier is a conventional supervised learner applied to a conventional feature set, and no claim of algorithmic novelty is made. The contribution lies in the comparison of two decision frameworks and in the analysis of the contexts in which each is defensible. A deliberately standard model supports that aim, because it attributes observed differences to the breadth of the evidentiary base rather than to algorithmic sophistication and makes the approach straightforward to reproduce elsewhere.

5.3 Analytic Procedures

For RQ1, agreement with terminal outcomes was examined across checkpoints in student-skill sequences to evaluate the consistency of rule-based and adaptive mastery decisions over repeated opportunities. For RQ2, a generalized estimating equation (GEE) model with robust standard errors was used to estimate the effects of sequence position, attempts, hint use, response time, skill, and template on rule-based mastery decisions. For RQ3, the adaptive model was compared with the traditional rule across skills and templates using agreement differences, examining both average improvement and negative cases. For RQ4, checkpoint-level error rates were compared to assess whether the adaptive model reduced context-related instability, particularly during the

early stages of learning. A separate negative-case analysis determined whether contexts in which the traditional rule outperformed the adaptive model reflected random variation or a systematic pattern.

6 Results

6.1 RQ1: Consistency Across Repeated Opportunities

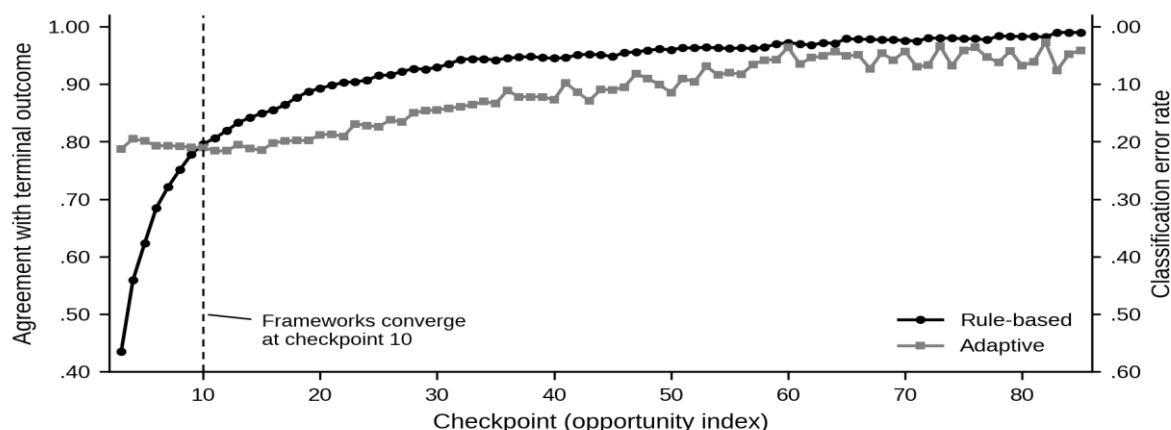
The adaptive model yielded substantially higher agreement with terminal outcomes during the early checkpoints of student-skill trajectories. At checkpoint 3, rule-based agreement was .435 against adaptive agreement of .787, an adaptive advantage of .353, and at checkpoints 4 and 5 the advantages were .246 and .178. The traditional criterion is therefore least stable precisely when mastery is still emerging and when instructional decisions are most consequential.

As Figure 1 shows, the two frameworks converged at checkpoint 10, after which the rule-based criterion showed the higher agreement rate among the smaller number of trajectories that continued that long. The adaptive advantage is therefore concentrated rather than uniform, and it is concentrated in the window where the decisions that determine whether a student advances or continues practicing are actually made.

Overall classification metrics showed the same pattern. The adaptive model achieved higher accuracy (.817 vs. .749) and higher F1 (.891 vs. .840), and showed good discrimination and calibration (AUC = .866, Brier score = .139), indicating that its estimated probabilities are suitable for threshold-based decision rules and not only as a binary verdict.

Figure 1

Agreement With Terminal Outcome and Classification Error Across Checkpoints



Note. Error rate is the complement of agreement, so both scales describe the same curves. The adaptive model showed higher agreement at the earliest checkpoints, where learner status was least certain. The two frameworks converged at checkpoint 10, after which the rule-based criterion showed higher agreement among the smaller number of extended trajectories. Checkpoints beyond 85 are omitted because fewer than 100 trajectories remained.

6.2 RQ2: Contextual Variation in Rule-Based Mastery Decisions

The GEE results showed that mastery decisions under the traditional rule varied significantly across sequence position, learner behavior, skills, and templates. Sequence position was positively associated with rule-based mastery ($b = 0.085$, $SE = 0.009$, $p < .001$), indicating that students were more likely to be classified as mastered later in the sequence. Hint use had a large negative effect ($b = -0.534$, $SE = 0.016$, $p < .001$), suggesting that reliance on hints substantially reduced the likelihood of mastery under the traditional rule. Attempt count also had a negative effect, although smaller in magnitude ($b = -0.008$, $SE = 0.003$, $p = .002$). Response time was not statistically significant once the other predictors were included ($b = 0.002$, $p = .901$).

Skill and template effects were substantial, as Figure 2 shows. Several skills were associated with significantly lower odds of rule-based mastery relative to the reference category, including Write Linear Equation from Graph ($b = -1.624$), Subtraction Whole Numbers ($b = -1.014$), and Equation Solving More Than Two Steps ($b = -0.737$). Template effects varied more widely still, ranging from $b = -1.559$ to $b = 1.086$, a spread of roughly 2.6 log-odds units between item templates nominally serving the same instructional function.

That spread is the substantive finding. The rule is stated in context-free terms, but its operating characteristics are not context-invariant. Whether three consecutive correct responses represents a demanding or a lenient bar depends heavily on which skill and which template a student happens to encounter, and on how much scaffolding that student has drawn on. Two learners of equivalent proficiency, practicing different templates within the same skill, face materially different probabilities of being declared a master.

6.3 RQ3: Comparative Consistency Across Skills and Templates

Across both skill-level and template-level analyses the adaptive model demonstrated a clear overall advantage (Table 1). Mean adaptive improvement in agreement was .120 across 77 skills and .095 across 349 templates, with positive median improvements in both cases. Weighted overall improvement remained positive after accounting for group size, indicating that the advantage was not an artifact of small and noisy contexts.

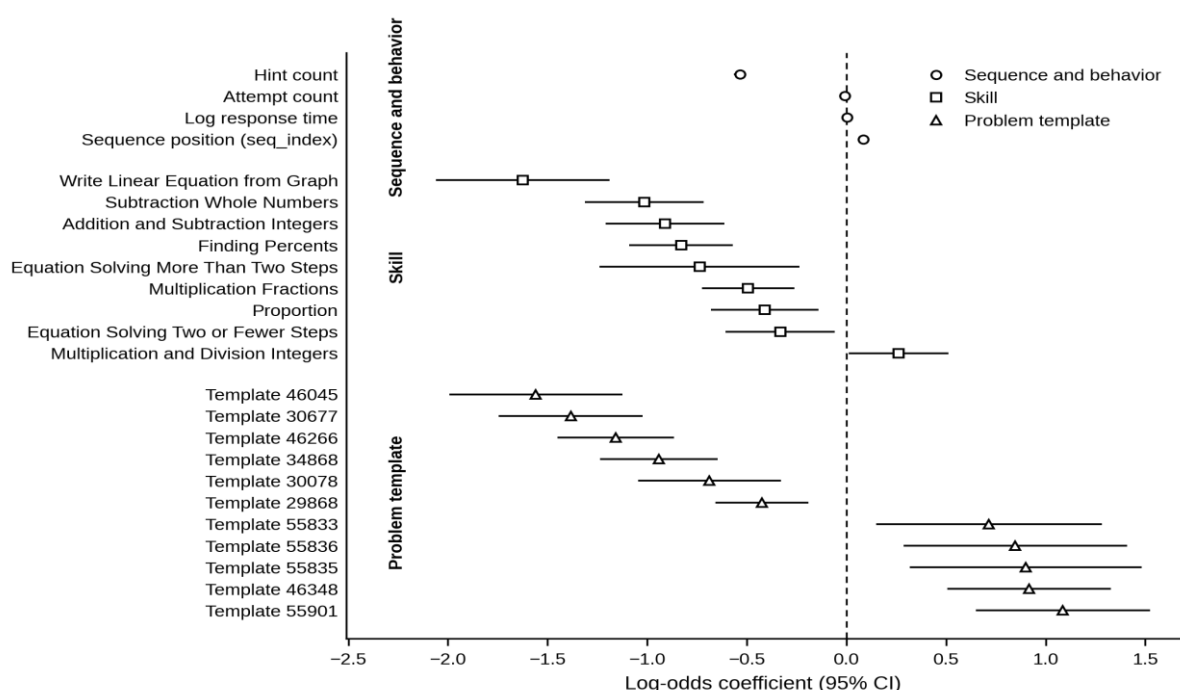
However, the adaptive advantage was not universal. Negative values appeared for 13 skills (16.9%) and 76 templates (21.8%), indicating cases in which the traditional rule outperformed the adaptive model. This is a non-trivial minority, and the pattern warranted closer inspection.

Table 1
Adaptive Improvement in Agreement Across Skills and Templates

Context	k	M	Mdn	Min	Max	Weighted M	Negative cases
Skills	77	.120	.118	−.320	.500	.074	13 (16.9%)
Templates	349	.095	.100	−.416	.488	.075	76 (21.8%)

Note. Adaptive improvement = adaptive agreement minus rule agreement; k = number of groups. Weighted M uses group sample size as weights. Negative cases are contexts in which the traditional rule outperformed the adaptive model.

Figure 2
Generalized Estimating Equation Predictors of Rule-Based Mastery



Note. Markers are log-odds coefficients and horizontal lines are 95% confidence intervals. Positive coefficients indicate increased odds of mastery under the traditional rule, whereas negative coefficients indicate decreased odds. Sequence position was positively associated with rule-based mastery, while hint use and repeated attempts were negatively associated with mastery. Substantial variation across skills and templates indicates that the three-correct-in-a-row criterion was context-sensitive rather than uniformly generalizable.

6.4 RQ4: Reduction of Context-Related Instability

The adaptive model substantially reduced error rates relative to the traditional rule at early checkpoints. Because agreement and error are complements, these values mirror the agreement results already shown in Figure 1 and are reported here for their instructional interpretation. At checkpoint 3, rule-based error was .565, compared with .213 for the adaptive model, an error reduction of .353. At checkpoint 4 the corresponding values were .441 and .195, and at checkpoint 5 they were .377 and .199.

The reduction diminished as sequences lengthened. Rule-based error fell steeply and monotonically, from .565 at checkpoint 3 to .279 at checkpoint 7 and lower thereafter, whereas adaptive error remained close to .20 throughout; the two frameworks converged at checkpoint 9 and the rule showed lower error beyond that point. The adaptive model is therefore not uniformly more accurate. It is front-loaded, reaching its operating level within the first few opportunities, while the rule improves only as consecutive evidence accumulates. It reduces misclassification precisely where learner status is most uncertain and where instructional decisions are least reversible.

6.5 Negative-Case Analysis

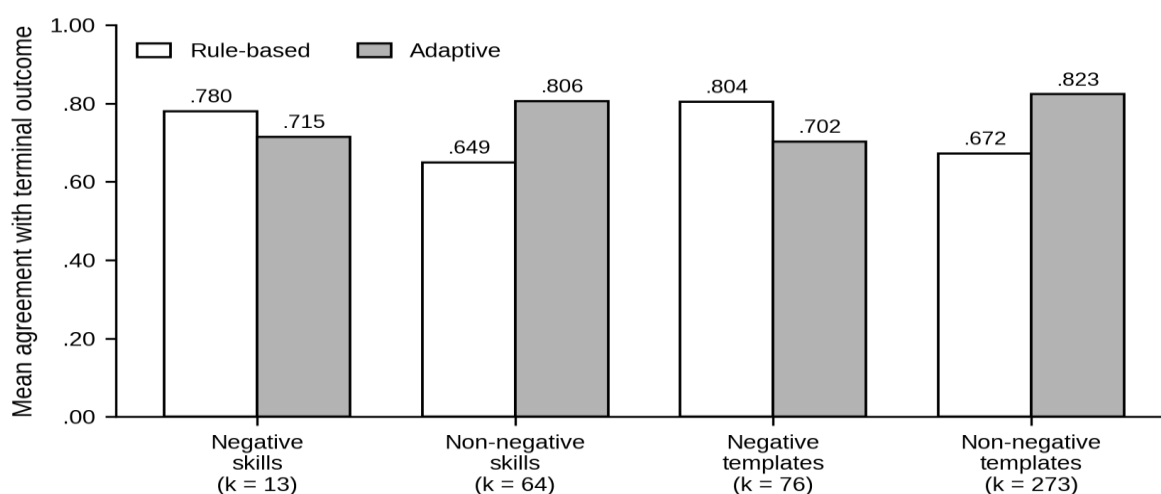
The negative-case analysis showed that the traditional rule did not outperform the adaptive model randomly. In negative skill cases, mean rule agreement was .780 compared with .715 for the adaptive model. In negative template cases, mean rule agreement was .804 compared with .702 for the adaptive model. In non-negative contexts the ordering reversed and the margin widened, with adaptive agreement of .806 across skills and .823

across templates against rule agreement of .649 and .672 (Figure 3).

The pattern is consistent across both groupings. Negative cases are contexts in which the rule was already achieving agreement of roughly .78 to .80, well above its average. Where the rule performs poorly the adaptive model improves substantially; where the rule already performs well it adds little and may cost a small amount. Negative cases therefore cluster in high-agreement contexts rather than representing arbitrary failures of the adaptive model.

Figure 3

Mean Agreement in Negative and Non-Negative Contexts



Note. Negative cases were defined as contexts in which adaptive agreement was lower than rule-based agreement; k = number of groups. In both skill-level and template-level negative cases the traditional rule showed relatively high mean agreement, indicating that negative cases were not random but clustered in contexts where the rule already performed well. This pattern supports a context-sensitive interpretation rather than unconditional replacement of the traditional mastery rule.

7 Discussion

This study examined whether a traditional rule-based mastery criterion remains defensible in an AI-supported formative assessment context. The three-correct-in-a-row rule proved operationally simple and sometimes effective, but not uniformly consistent or context-invariant. Its performance varied across learning stages, learner support patterns, skills, and templates, and it was weakest during the early sequence stages, where mastery decisions are most uncertain and instructional consequences most significant (Beck & Gong, 2013).

By contrast, the adaptive model provided stronger early-stage classification and reduced early-stage error, suggesting that mastery

decisions benefit from a broader evidentiary base than simple consecutive correctness (Corbett & Anderson, 1995; Pavlik, Cen, & Koedinger, 2009). From a measurement perspective this supports movement toward adaptive, data-informed mastery frameworks.

At the same time, the negative-case results show that the adaptive model is not universally superior. Where the traditional rule already has high agreement, particularly in tightly structured problem settings, the adaptive model may add limited value or even perform slightly worse. The results therefore support context-sensitive implementation rather than blanket replacement of rule-based mastery criteria (Käser, Klingler, & Gross, 2016).

7.1 Why Performance Varies Across Contexts

This variation has a straightforward explanation, which clarifies when each framework should be preferred. The traditional rule is a deterministic function of three binary observations. Where item difficulty within a skill is homogeneous, templates are tightly structured, and correct responses rarely arise from guessing, that function is close to sufficient, and little residual variance remains for additional evidence to explain. The high rule agreement in negative-case contexts, .780 across skills and .804 across templates, is consistent with this. In such settings a more flexible model adds estimation variance without a corresponding gain in accuracy.

Where contexts are heterogeneous, three consecutive correct responses are a noisier indicator. Rule agreement falls to roughly .65, and the additional indicators available to the adaptive model recover substantial information. The GEE finding that hint count carries a large negative coefficient describes the same phenomenon from the rule's perspective: hint reliance changes what a streak of correct responses means, and the rule provides no channel for that information. The adaptive model gains most where correctness alone is least informative, which is also where the instructional risk of a wrong decision is highest.

7.2 Practical Significance of the Observed Improvements

The practical significance of these improvements can also be expressed in the decisions they change rather than in classification metrics alone. At checkpoint 3, the error reduction of .353 applied to 25,500 observed checkpoints corresponds to approximately 9,002 mastery decisions that the adaptive model resolved correctly, and the traditional rule did not. Across checkpoints 3 through 7, which together comprise 94,786 checkpoint decisions, the cumulative figure is on the order of 20,400 decisions, or roughly 21.5% of all decisions made in that window. Because checkpoints are nested within students and sequences, these counts describe the volume of decision changes that a system operator would observe rather than a count of independent events. The order of magnitude, however, is the quantity relevant to an operational decision.

Each such decision carries a specific instructional consequence. A false mastery classification advances a student who has not yet consolidated the skill, withdrawing the practice

that would have consolidated it and displacing the resulting deficit into later content that depends on it. A false non-mastery classification holds a student in continued practice on a skill already acquired, consuming instructional time and, in prolonged cases, motivation (Beck & Gong, 2013). In a Skill Builder sequence, where the mastery decision determines whether a further item is served, these consequences are immediate rather than hypothetical.

The two errors are not equally costly, and their relative cost is a property of the instructional setting rather than of the model. In a remedial or prerequisite-heavy sequence, premature advancement is the more damaging error, and the system should require stronger evidence before declaring mastery. Where student time is the binding constraint, unnecessary continued practice is the greater cost. An adaptive framework accommodates this directly, because it yields a calibrated probability rather than a binary verdict, so the classification threshold becomes an explicit policy setting that an institution can align with the stakes of the sequence. The traditional rule offers only the coarser alternative of lengthening the required streak, which shifts both error types at once and without calibration. Operational reporting should therefore decompose classification error into premature advancement and unnecessary continued practice rather than reporting overall accuracy alone.

8 Implications and Conclusion

The study has three main implications. First, formative systems should treat mastery decisions as measurement claims requiring empirical validation rather than operational settings (Messick, 1989). Second, fixed thresholds such as three-correct-in-a-row should be reevaluated in light of the richer evidence now available in digital learning systems. Third, AI-supported formative assessment is most defensible when used to improve the consistency and instructional quality of classification decisions rather than to automate legacy heuristics (Black & Wiliam, 1998).

These implications point toward a modest implementation path rather than wholesale redesign. Because negative cases cluster in contexts where the traditional rule already achieves high agreement, a system can audit its existing rule-based decisions skill by skill and template by template, retain the rule where

historical agreement is high and the context is structurally regular, and deploy the adaptive criterion where agreement is low or variable. Such an audit requires only data the system already generates, and it preserves the transparency of the simple rule where that rule is defensible, which matters when mastery decisions must be explained to teachers, students, and families. Treating the mastery threshold as a claim requiring evidence in each context of use is itself a validity practice in Messick's (1989) sense.

Several boundaries on these conclusions should be stated plainly. Agreement was measured against the terminal outcome of the student-skill sequence rather than against an external posttest, so a framework can be consistent with the system's own resolution without that resolution indicating durable learning. The dataset is also historical, so the specific skill and template effects should not be assumed to transfer to a contemporary platform, although the argument for per-context validation is intended to. Finally, the study evaluates classification consistency rather than instructional impact.

Overall, the three-correct-in-a-row rule is not sufficiently robust as a universal mastery decision rule for contemporary formative assessment systems. An adaptive, measurement-informed model provides more consistent and instructionally defensible decisions during the early stages of learning, where instructional stakes are highest and where roughly a fifth of all early mastery decisions would be resolved differently. At the same time, the negative-case evidence indicates that rule-based mastery may still be appropriate in selected high-regularity settings that can be identified empirically in advance. The most defensible path forward is therefore not replacement but the development of context-sensitive, evidence-based mastery frameworks for AI-integrated formative assessment.

Data Availability

The ASSISTments "Skill Builder" 2009–2010 corrected dataset is publicly available at <https://sites.google.com/site/assistmentsdata/home/2009-2010-assistment-data/skill-builder-data-2009-2010>.

References

Beck, J. E., & Gong, Y. (2013). Wheel-spinning: Students who fail to master a skill. In H. C. Lane,

- K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial intelligence in education (AIED 2013)* (pp. 431–440). Springer.
https://doi.org/10.1007/978-3-642-39112-5_44
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74.
<https://doi.org/10.1080/0969595980050102>
- Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4), 253–278.
<https://doi.org/10.1007/BF01099821>
- Feng, M., Heffernan, N. T., & Koedinger, K. R. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction*, 19(3), 243–266. <https://doi.org/10.1007/s11257-009-9063-7>
- Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4), 470–497.
<https://doi.org/10.1007/s40593-014-0024-x>
- Käser, T., Klingler, S., & Gross, M. (2016). When to stop? Towards universal instructional policies. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (LAK '16)* (pp. 289–298). Association for Computing Machinery.
<https://doi.org/10.1145/2883851.2883961>
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.
- Pardos, Z. A., Bergner, Y., Seaton, D. T., & Pritchard, D. E. (2013). Adapting Bayesian knowledge tracing to a massive open online course in edX. In *Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013)* (pp. 137–144). International Educational Data Mining Society.
- Pavlik, P. I., Cen, H., & Koedinger, K. R. (2009). Performance factors analysis: A new alternative to knowledge tracing. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED 2009)* (pp. 531–538). IOS Press.