

# Reliability Estimation Methods for Bayesian Knowledge Tracing Mastery Classifications

Samuel D. Ihlenfeldt and Stephen E. Fancsali

Carnegie Learning, Inc.

Pittsburgh, PA, USA

{sihlenfeldt, sfancsali}@carnegielearning.com

## Abstract

We evaluate the psychometric reliability of Bayesian Knowledge Tracing for skill mastery classification. Using large-scale intelligent tutoring system data, we compare two simulation-based approaches for estimating classification consistency as reliability. Results support both approaches, show reliability increases with response-sequence length, and highlight the necessity of adjusting for chance agreement.

## 1 Introduction

Knowledge tracing (KT) generally refers to a family of computational methods designed to model a student’s evolving knowledge state as they interact with a sequence of learning materials. KT methods are often used to predict students’ future performance and tailor instruction within learning applications like intelligent tutoring systems (ITSs; Abdelrahman et al., 2023). One of the earliest and most widely employed KT approaches is Bayesian Knowledge Tracing (BKT; Corbett & Anderson, 1994), a Hidden Markov Model that estimates a binary mastery state for a knowledge component (or skill) on the basis of four parameters: prior knowledge, probability of learning at a particular practice opportunity, and rates of “slipping” and “guessing” (Fancsali et al., 2013a; Ritter et al., 2007; Yudelson et al., 2013). Presumably, the parsimony of the model and interpretability of the parameters are major factors driving its continued dominance in operational settings over more recent models like Deep Knowledge Tracing (DKT) (Piech et al., 2015; Badrinath et al., 2021), which use opaque neural networks to learn complex dependencies across student learning sequences (Abdelrahman et al., 2023; Krivich et al., 2025).

Although BKT has been explored extensively within the educational data mining literature (Abdelrahman et al., 2023), there is less literature connecting KT approaches to traditional psychomet-

rics (Khajah et al., 2014; Shimada & Okada, 2025). Since BKT-based assessments are used to make diagnostic classifications with educational consequences for students, we might also consider applying psychometric standards to KT methods like those expected of traditional assessments. The present study focuses on one such standard: reliability, a fundamental requirement for any operational measurement model (American Educational Research Association et al., 2014). In measurement contexts with categorical outcomes, such as the mastery judgments derived from BKT, reliability is operationalized as classification consistency: the agreement of binary mastery classifications across hypothetical independent test replications (Gierl et al., 2009; Lee & Harris, 2025). Unlike classical measurement models, though, KT assumes that students learn through interaction with the instrument itself (e.g., while interacting with learning materials in an ITS), necessitating methods that incorporate student response sequences into those hypothetical replications.

Only a handful of studies have examined reliability in the context of BKT (Cui et al., 2019; Shimada & Okada, 2025). Indeed, prior researchers have criticized the data mining literature for overemphasizing model fit or prediction performance (Krivich et al., 2025) while paying less, if any, attention to the psychometric properties of the underlying models (Shimada & Okada, 2025). Although model fit and predictive performance are important components for trustworthy models, prior work suggests a high area under the curve (AUC) has an inconsistent relationship with high reliability estimates in BKT contexts (Shimada & Okada, 2025). Consequently, a model can appear to be of high quality by the standard of prediction performance while producing classifications that are not reliable by psychometric standards.

These observations point to the need to build on existing literature (Cui et al., 2019; Shimada &

Okada, 2025) to better understand and quantify the reliability of BKT models. The interpretable parameter structure of BKT makes it possible to observe and reason about the aspects of the model that may compromise reliability. In the present research, we specifically investigate the consistency of the mastery decision itself. Two methods for establishing the classification consistency of BKT models are examined: one extant (Almond et al., 2015) and one novel. The empirical viability of both methods is subsequently demonstrated through their application to a large-scale dataset derived from a widely deployed ITS.

## 2 Classification Consistency

Although a robust body of literature addresses classification consistency within contexts such as item response theory (IRT), these approaches generally evaluate reliability by setting a cutscore on a continuous latent trait distribution (Lee, 2010; Rudner, 2000). In contrast, BKT fundamentally assumes a categorical latent state, explicitly modeling students as either “masters” or “non-masters” of specific skills. Because of this, the psychometric evaluation of BKT naturally aligns with the methodologies developed within the cognitive diagnostic assessment (CDA) literature.

Over the last two decades, many researchers have designed methods for quantifying the reliability of CDAs (Cui et al., 2012; Gierl et al., 2009; Johnson & Sinharay, 2020; Templin & Bradshaw, 2013; Templin & Henson, 2009; Wang et al., 2015; Yamaguchi & Templin, 2022). Recently, some researchers have advanced classification consistency estimation by simulating hypothetical test replications directly from the empirical posterior mastery estimates of the examinees (Thompson et al., 2019, 2023). Although these CDA posterior-simulation approaches ground reliability estimates in observed data, they remain firmly anchored in the assumption that students are not learning as they interact with the instrument.

### 2.1 Reliability of BKT Classifications

A few recent efforts have begun to address the psychometric properties of BKT models specifically, including Shimada and Okada (2025), who directly adapted their approach from the CDA literature (Templin & Bradshaw, 2013). Their method produced both an overall and time-dependent reliability estimate of the continuous probability of

knowledge acquisition in BKT models by approximating the tetrachoric correlation of a  $2 \times 2$  matrix that cross-tabulated a student’s joint probabilities of mastery and non-mastery across two hypothetical testing occasions. These metrics work for contexts in which probability of mastery is important, such as for progress monitoring, but instructional decisions in ITSs are often driven by the binary classification decisions for whether a particular skill is judged to be mastered (e.g., by the probability estimate reaching a threshold such as 0.95; Corbett & Anderson, 1994). We aim to bolster the psychometric foundations of KT methods by estimating reliability for a metric used to make specific educational inferences (Johnson & Sinharay, 2018; Wang et al., 2015), in this case the binary classification of skill mastery rather than the underlying probability of having attained such mastery.

Almond et al. (2015) pioneered the evaluation of classification consistency from a Bayesian perspective. Specifically, they estimated an *accuracy matrix* that compared “true” classifications from generated skill profiles to the classifications from simulated response strings. This matrix was then multiplied by the transpose of a simulated parallel form’s accuracy matrix, yielding a single *consistency matrix* (which could also be estimated by directly comparing the parallel forms). While their framework was designed to handle complex, multi-category proficiency profiles, they noted that it simplifies to a basic binary setup. Based on such a setup, this simulation methodology was practically applied by Cui et al. (2019), who strongly endorsed the *Almond Method* as a framework to estimate the reliability of a formative game-based assessment with an underlying BKT structure.

It is crucial to note that the Almond Method is fundamentally a model-based approach, so it evaluates the reliability of the theoretical model itself. During the simulation process, the model’s fitted parameters are treated as the absolute truth and the empirical student responses are ignored in favor of generating idealized response patterns. Further, a critical difference between these implementations and the present analysis lies in the test structure: both Almond et al. (2015) and Cui et al. (2019) assumed a fixed number of tasks for each skill across all students. However, the ITS in the present study adapts to student learning, so the method must account for dynamic test lengths based on when students achieve mastery.

### 3 Estimation Techniques

#### 3.1 Almond Method

The simulation framework introduced by Almond et al. (2015) provides a model-based approach to estimating the classification consistency for a skill. This approach treats the global parameters of the model as a data-generating mechanism, simulating response sequences for a theoretical population of examinees that align with the model's parameters. By sampling true proficiency states from generative priors and projecting simulated evidence through the model's parameters, the consistency with which a standard BKT engine recovers the true mastery state for a given skill can be quantified.

##### 3.1.1 Sampling a Transition Point

For a single skill evaluated across a sequence of  $N$  items, each simulee is assigned a true transition point  $k_{\text{true}}$ , which represents the exact item step at which they transition from non-mastery to mastery. In this framework,  $N$  is established a priori and is constant across simulees.  $k_{\text{true}}$  is sampled directly from the geometric prior distribution governed by the population parameters, initial knowledge,  $P(L_0)$ , and learning,  $P(T)$ .

##### 3.1.2 Simulating Response Vectors

Using  $k_{\text{true}}$ , two parallel, independent response vectors (Run A and Run B) are simulated for each student. Following the standard BKT generative process, responses prior to the transition ( $j < k_{\text{true}}$ ) are generated using the guessing probability ( $P(Y_j = 1) = P(G)$ ), while responses at or after the transition ( $j \geq k_{\text{true}}$ ) are generated using the non-slip probability ( $P(Y_j = 1) = 1 - P(S)$ ). Both parallel response vectors are then processed independently through the standard BKT filtering equations to calculate the final probability of mastery,  $P(L_t)$ . Consistent with the operational definition of mastery, a student is formally classified as a master once the updated probability, evaluated after every item response, reaches or exceeds 0.95; this threshold yields the final binary mastery classifications  $\hat{\alpha}^{(A)}$  and  $\hat{\alpha}^{(B)}$ .

#### 3.2 Posterior Predictive Classification Consistency

To evaluate the classification consistency of a sample of responses, an empirical simulation framework called posterior predictive classification consistency (PPCC), rooted in the logic of posterior

predictive model checking (Thompson et al., 2019, 2023), is proposed. Unlike the Almond method, which draws hypothetical examinees from population priors (Almond et al., 2015), this method additionally uses the empirical student response strings to generate data. Each examinee's observed response string is used to derive a unique posterior probability distribution,  $K$ , for their exact moment of mastery transition. We then use this posterior distribution to test the consistency of the classification via parallel Monte Carlo replications.

##### 3.2.1 Empirical Posterior

Let  $N$  be the number of learning opportunities for a student, and  $\mathbf{Y}$  be the vector of observed binary responses. By applying Bayes' theorem, the prior probability of the transition point is combined with the conditional likelihood of the student's actual response string (given the theoretical model parameters). This allows the empirical posterior probability distribution of the transition point  $k$  to be calculated, conditioned entirely on  $\mathbf{Y}$ :

$$P(K = k | \mathbf{Y}) = \frac{P(\mathbf{Y} | K = k) P(K = k)}{\sum_{m \in \Omega_K} P(\mathbf{Y} | K = m) P(K = m)}, \quad (1)$$

where  $\Omega_K = \{0, 1, \dots, N, > N\}$  represents the complete sample space of potential transition points. This equation yields a unique probability mass function for each student based on their individual responses.

##### 3.2.2 Monte Carlo Simulation

A simulated transition point  $k_{\text{true}}$  is randomly sampled from the posterior distribution for each student:

$$k_{\text{true}} \sim P(K | \mathbf{Y}) \quad (2)$$

Using  $k_{\text{true}}$ , the model simulates two parallel, independent sets of item responses, Run A ( $\mathbf{Y}^{(A)}$ ) and Run B ( $\mathbf{Y}^{(B)}$ ), injecting noise with  $P(G)$  and  $P(S)$ . For steps before the transition ( $j < k_{\text{true}}$ ), the response is generated using the guessing probability ( $P(Y_j = 1) = P(G)$ ). For steps at or after the transition ( $j \geq k_{\text{true}}$ ), the response is generated using the non-slip probability ( $P(Y_j = 1) = 1 - P(S)$ ). Unique to this method, the length of these simulated response vectors is identical to the length  $N$  of a specific student's observed response string. These parallel response vectors are then processed back through the standard BKT updating equations to generate two independent final mastery classifications,  $\hat{\alpha}^{(A)}$  and  $\hat{\alpha}^{(B)}$ , based on the operational 0.95 cutoff.

### 3.3 Estimating Classification Consistency

For each simulee, a  $2 \times 2$  probability matrix is constructed across  $\mathbf{Y}^{(A)}$  and  $\mathbf{Y}^{(B)}$  for each pair of possible  $\hat{\alpha}^{(A)}$  and  $\hat{\alpha}^{(B)}$ . Each cell is then averaged across all simulees to construct a marginal joint probability matrix,  $\mathbf{P}$ . To calculate consistency,  $\mathbf{P}$  is then passed directly to the formulas for non-chance and chance-adjusted agreement metrics (Johnson & Sinharay, 2018).

## 4 Empirical Demonstration

To evaluate the viability of the proposed metrics, this section presents an empirical demonstration using student response data from an operational ITS.

### 4.1 Data

Data for this analysis were drawn from a widely studied ITS that employs BKT (Abdelrahman et al., 2023). The robust dataset includes problem sets from a 7th-grade math curriculum collected across 1,204 schools.

Model BKT parameters were generated using *hmm-scalable* (Yudelso, 2017) and all other analyses were conducted in *R* (R Core Team, 2026). The students were divided into a 75% training set and a 25% evaluation set independently for each skill. The guessing parameter was capped at 0.25 to ensure that models were interpretable; without this constraint, the EM algorithm frequently converged to solutions with a degenerate guessing parameter. This phenomenon is well-documented in extant BKT literature, violating the assumption that knowledge leads to higher performance than guessing (Baker et al., 2008). The final dataset, consisting of the 25% evaluation set, included 66 distinct skills distributed across six problem sets, with each problem set contributing between 5 and 15 skills (Table 1). Figure 1 depicts the distribution of parameters for all 66 skills. Notably, the distribution of guessing parameters is heavily left skewed because of this cap, and 39 of the 66 skills had  $P(G)$  equal to 0.25.

### 4.2 Evaluation Metrics

In a critique of CDA methods, Johnson & Sinharay (2018) demonstrated that raw classification consistency can be artificially inflated if a classification has high or low prevalence in the target population. To combat this, they argued that attribute-level reliability must be evaluated using chance-

| PS | Skills | Students | Practice Depth | % Mastery |
|----|--------|----------|----------------|-----------|
|    |        | 14,441–  |                | 95.09–    |
| 1  | 5      | 14,466   | 6–7            | 99.98     |
|    |        | 11,421–  |                | 80.63–    |
| 2  | 7      | 11,442   | 6–36           | 100       |
|    |        | 7,456–   |                | 82.77–    |
| 3  | 9      | 7,476    | 3–14           | 100       |
|    |        | 6,264–   |                | 75.31–    |
| 4  | 15     | 6,272    | 2–7            | 99.97     |
|    |        | 12,702–  |                | 80.65–    |
| 5  | 15     | 12,748   | 3–8            | 100       |
|    |        | 13,816–  |                | 64.53–    |
| 6  | 15     | 13,867   | 2–10           | 100       |

Table 1: Summary of the six problem sets. Practice depth is the median number of opportunities seen by students for a skill.

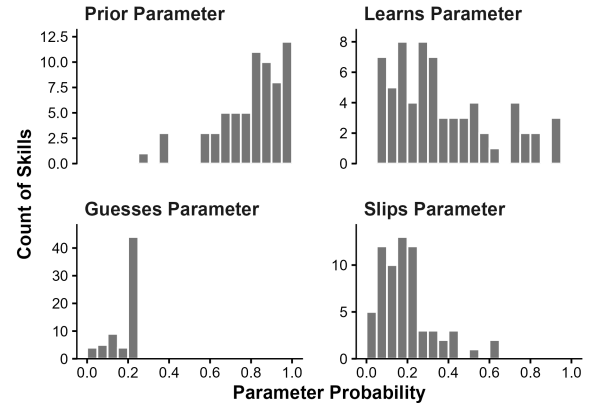


Figure 1: Distribution of fitted BKT parameters across all 66 skills.

adjusted agreement metrics to accurately reflect an assessment’s true diagnostic value independent of population skew. BKT models assume that every student will achieve mastery with enough time, resulting in near-perfect raw agreement between simulees. Given the relatively high rate of mastery across every skill in the study, four distinct agreement metrics were adopted to evaluate classification consistency. The first is raw agreement, which represents the absolute proportion of matching classifications between parallel simulations. To adjust for population skew, we also included the three chance-corrected methods recommended by Johnson & Sinharay (2018). Cohen’s  $\kappa$  evaluates the degree of agreement beyond what would be expected by a purely random baseline. Youden’s  $J$  evaluates the overall diagnostic effectiveness of the model by capturing the proportion of informed classifications beyond chance. It weights

the model's ability to correctly identify both true masters and true non-masters equally, calculated as  $J = \text{sensitivity} + \text{specificity} - 1$ . Finally, Goodman and Kruskal's  $\lambda$  measures the proportional reduction in classification error when using the model compared to a naive modal baseline (in this case, blindly assuming every student is a master). Because the base rate of mastery is so high, this naive baseline is already highly accurate, meaning the model faces a strict penalty for any misclassifications. Unlike  $J$  and  $\kappa$ , which are bounded from  $-1$  to  $1$ ,  $\lambda$  can become unboundedly negative when a model underperforms the modal baseline in highly skewed distributions.

### 4.3 Results

To calculate the Almond metric and compare it to PPCC estimates, the median response length was used as the basis for generating simulees (e.g., for a skill with a median response length of 5 items, 1,000 pairs of simulees were generated with 5 responses). As shown in Figure 2, raw agreement yielded systematically higher overall reliability estimates, with both the Almond and PPCC methods heavily concentrated near the 1.0 ceiling and exhibiting minimal variance.<sup>1</sup> This high raw agreement is driven by the high base rate of mastery in the sample. However, when controlling for chance agreement and base rates, reliability estimates dropped substantially across the board. Comparing the two estimation approaches, the Almond method consistently produced higher median reliability coefficients and wider upper ranges than the PPCC method across all three chance-adjusted metrics. Among these adjusted metrics,  $\kappa$  and  $J$  displayed nearly identical distributions within each respective method. Conversely,  $\lambda$  yielded the lowest overall estimates and the widest variance, with the Almond median at zero and the majority of the PPCC distribution falling into the negative range. Because  $\lambda$  gives the model zero credit for predicting the dominant “master” class, it reflects only how well the algorithm isolates the remaining non-masters. A Bonferroni-corrected paired Wilcoxon test indicates that raw agreement did not significantly differ between frameworks ( $p_{\text{adj}} = 0.246$ ), but the Almond method yielded higher reliability estimates than PPCC across all three chance-adjusted metrics ( $p_{\text{adj}} < 0.001$ ).

<sup>1</sup>The plots in this section summarize the primary findings; a more detailed table containing skill-specific results is available upon request.

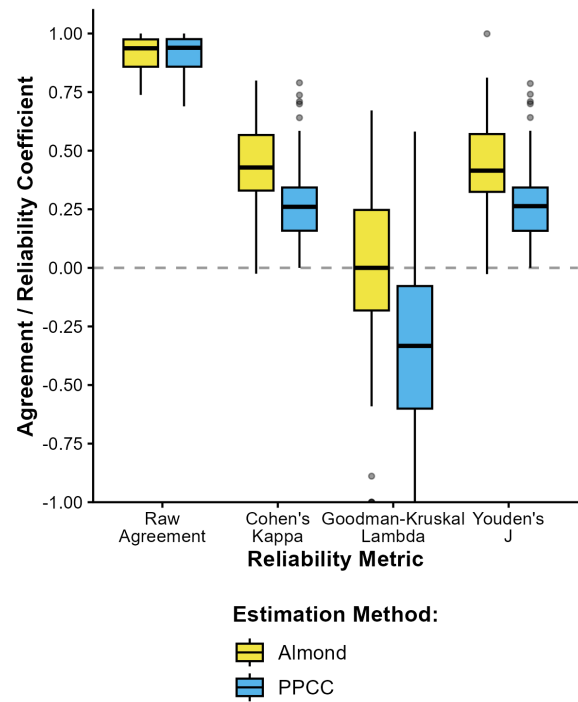


Figure 2: Distribution of attribute-level reliability metrics.

To evaluate reliability, increasing  $N$  (Figure 3) was operationalized differently for the two measurement frameworks. For the Almond method, each point along the x-axis represents an independent simulation of 1,000 simulees evaluated at exactly  $N$  items. For the PPCC method, the calculations at step  $N$  restricted the sample to the empirical subset of students who completed at least that many items, and treated the decision consistency as if students were to stop at that point. As  $N$  increases, both methods experience data attrition that impacts the reliability estimates. In the PPCC method, the underlying sample size inherently shrinks as students with shorter practice histories are filtered out. In the Almond method, because the BKT framework treats mastery as an absorbing state (Corbett & Anderson, 1994), longer response sequences force the simulated parallel runs to eventually converge on uniform mastery ( $\hat{\alpha}^{(A)} = \hat{\alpha}^{(B)} = 1$ ). To account for the mathematical artifacts generated when individual skill models reached this absolute consensus, these instances were excluded from the calculations of the chance-adjusted metrics ( $\kappa$ ,  $\lambda$ , and  $J$ ). In these instances of perfect agreement, classification variance drops to zero, rendering the denominators of the adjusted metrics mathematically undefined. Consequently, 837 observations (29%, almost en-

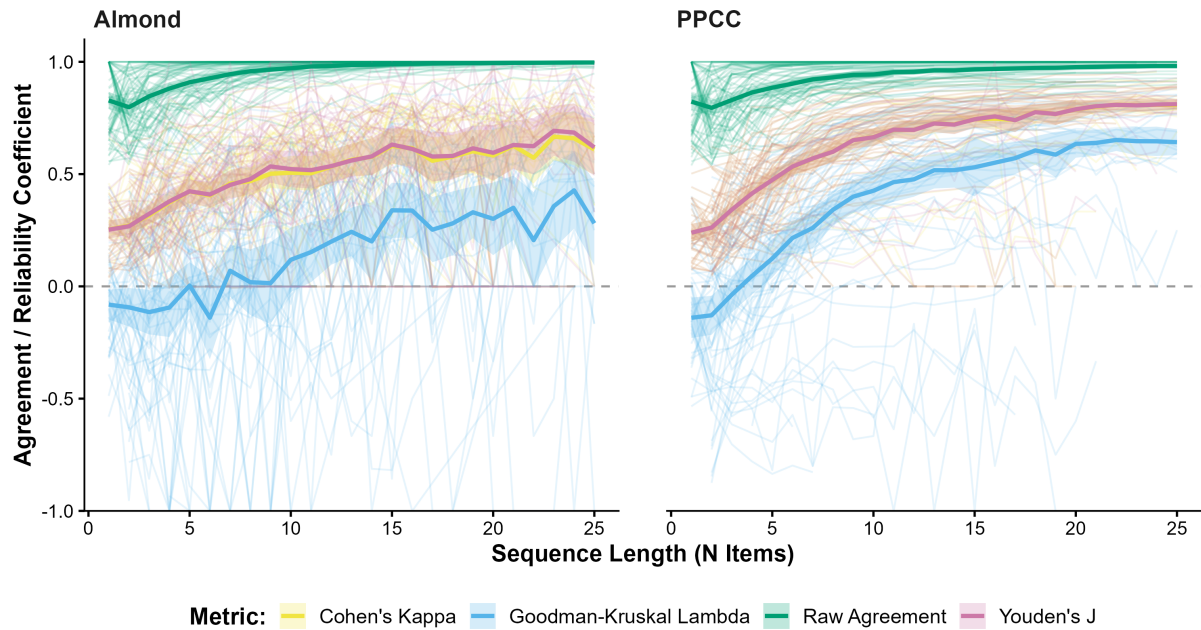


Figure 3: Reliability metrics against sequence length. For the Almond approach, 1,000 simulees were generated per skill.

tirely occurring at high values of  $N$ ) were coded as missing to prevent artificial downward spikes in the reported averages. This combined data attrition is visually reflected in the widening confidence ribbons at longer sequence lengths.

Referring again to Figure 3, all evaluated metrics exhibited a general upward trajectory as sequence length increased. Raw agreement rapidly approached the 1.0 ceiling within the first five items for both methods and remained consistently flat.  $\kappa$  and  $J$  tracked almost identically, starting lower but climbing steadily before beginning to plateau at higher values of  $N$ .  $\lambda$  consistently yielded the lowest estimates, starting near or below zero for the shortest sequences before gradually trending upward. Although the theoretical Almond curves showed volatility at high values of  $N$ , the empirical PPCC lines remained stable.

## 5 Discussion

Although work in educational data mining often relies on predictive fit to evaluate models, trustworthy measurement also requires explicit evidence of reliability; therefore, this paper set out to establish whether the classification consistency of BKT models can be estimated using psychometrically defensible frameworks. Two methods were considered: the established, model-based Almond method (Almond et al., 2015; Cui et al., 2019), and a novel,

data-driven method, PPCC, which simulates parallel replications from each student’s posterior predictive distribution over their moment of mastery transition (Thompson et al., 2019, 2023). Taken together, this paper supports the use of both methods, with the PPCC method offering a complement to the theoretical Almond approach; in particular, the empirical grounding of PPCC is crucial when moving from model evaluation to operational psychometric validation, as it evaluates classification consistency informed by observed student responses rather than relying entirely on theoretical model-driven response patterns. A post hoc analysis of AUC (the model’s item-level predictive discrimination) further reinforced the need for psychometric evidence beyond model fit, exhibiting only moderate correlations with the chance-adjusted metrics of both the Almond ( $r = .48 - .65$ ) and PPCC ( $r = .34 - .48$ ) frameworks.

For each agreement metric, the PPCC approach produced estimates similar to, albeit slightly lower than, those of the established Almond method, lending credibility to its use while reflecting the additional noise present in real student responses. Notably, both the Almond method and PPCC failed in different but important ways at higher values of  $N$ . For the Almond method, skills with sufficiently high  $P(T)$  values yielded 100% mastery. Because the simulated BKT models treat mastery as an ab-

sorbing state, given a sufficiently long response sequence, the parallel classification decisions by definition converge to uniform mastery for all simulated students. This feature of BKT yielded perfect raw agreement but mathematically undefined chance-adjusted indices, typically at a higher  $N$ . For the PPCC method, the empirical sample experienced attrition bias because stronger students achieved mastery early. However, the stability of PPCC at these higher intervals suggests that for the specific subpopulation of students who have the highest number of learning opportunities, the underlying BKT model functions as a highly reliable assessment instrument. Both results together suggest requiring a minimum number of items with a skill may be requisite to attaining an acceptable level of reliability.

Although not the primary intention of this study, these results strongly support the conclusion of Johnson & Sinharay (2018) that chance correction is not optional. Solely observing raw agreement yielded a conclusion that was not consistent with the other agreement metrics. The values of Cohen's  $\kappa$  and Youden's  $J$  were nearly identical to each other, but still far lower than raw agreement across all  $N$ . Mathematically, whenever a  $2 \times 2$  matrix achieves perfect marginal symmetry, the algebraic formulas for Cohen's  $\kappa$  and Youden's  $J$  collapse into the exact same equation, which occurred in the present study. In contrast, Goodman and Kruskal's  $\lambda$  is a modal metric: it asks how much the model improves over a blind baseline strategy. In the present research, that baseline already correctly classifies the vast majority of students as masters. Consequently, when evaluating a population with high skew, there is added value in using  $\lambda$  because it strictly focuses on the model's ability to determine mastery beyond a naive modal baseline.

## 5.1 Future Research

This study demonstrates the viability of both the PPCC and Almond frameworks; however, several key relationships between underlying BKT model features and these reliability estimation methods remain unresolved. Simulation studies are a practical path forward for understanding the nature of these relationships in both instances. Because the Almond method is entirely theoretical, its utility depends heavily on the specific configuration of the estimated BKT parameters. Future work should evaluate the direct impact of manipulating model parameters or assumptions on the four

consistency metrics, as well as the 0.95 mastery threshold (Fancsali et al., 2013b). Beyond parameter sensitivity, future research should explore the existing methodology proposed by Shimada and Okada (2025). Although their framework evaluates reliability at the latent state distribution level rather than the binary classification that is the focus of the present research, there remains value in analyzing the reliability of the underlying distribution of mastery as a complementary validation tool. Lastly, under operational conditions, there are a myriad of reasons why a student might see additional items (up to  $N_{\max}$  items), so there is some concern that the present methods truncate a response string and compare mastery decisions prematurely in the PPCC replications. More accurately understanding classification consistency may necessitate a framework that extends the posterior distribution using the model's learning parameter, beyond the exact  $N$  items seen by a student and up to  $N_{\max}$  for a given skill.

As KT models continue to drive ITSs, establishing robust methodologies for quantifying their reliability is essential for innovative use cases (e.g., as an alternative to traditional high-stakes assessments). The present study demonstrates that both theoretical model-based and empirical posterior-predictive simulations offer distinct insights into classification consistency. Crucially, establishing these foundational baselines for standard BKT parameters is a necessary prerequisite before evaluating how individual skill-level reliability impacts the consistency of macro-level ITS decisions, such as advancing a student through mixed-skill problem sets or curricular modules. Furthermore, establishing this methodological groundwork is essential before scaling up psychometric evaluation to the more complex-yet-opaque DKT architectures that dominate current research.

## Acknowledgments

This work was supported, in whole or in part by the Gates Foundation [INV-072481]. The conclusions and opinions expressed in this work are those of the author(s) alone and shall not be attributed to the Foundation. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 License has already been assigned to the Author Accepted Manuscript version that might arise from this submission.

## References

- Abdelrahman, G., Wang, Q., & Nunes, B. P. (2023). Knowledge tracing: A survey. *ACM Computing Surveys*, 55(11):1–37.
- Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). *Bayesian networks in educational assessment*. Springer.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Badrinath, A., Wang, F., & Pardos, Z. (2021). pyBKT: An accessible Python library of Bayesian knowledge tracing models.
- Baker, R. S. J., Corbett, A. T., & Aleven, V. (2008). More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian knowledge tracing. In *Intelligent Tutoring Systems*, pages 406–415. Springer.
- Corbett, A. T. & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278.
- Cui, Y., Chu, M.-W., & Chen, F. (2019). Analyzing student process data in game-based assessments with Bayesian knowledge tracing and dynamic Bayesian networks. *Journal of Educational Data Mining*, 11(1):80–100.
- Cui, Y., Gierl, M. J., & Chang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, 49(1):19–38.
- Fancsali, S. E., Nixon, T., & Ritter, S. (2013a). Optimal and worst-case performance of mastery learning assessment with Bayesian knowledge tracing. In *Proceedings of the 6th International Conference on Educational Data Mining*, pages 35–42. International Educational Data Mining Society.
- Fancsali, S. E., Nixon, T., Vuong, A., & Ritter, S. (2013b). Simulated students, mastery learning, and improved learning curves for real-world cognitive tutors. In *AIED 2013 Workshops Proceedings Volume 4*, page 11.
- Gierl, M. J., Cui, Y., & Zhou, J. (2009). Reliability and attribute-based scoring in cognitive diagnostic assessment. *Journal of Educational Measurement*, 46(3):293–313.
- 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.
- Johnson, M. S. & Sinharay, S. (2020). The reliability of the posterior probability of skill attainment in diagnostic classification models. *Journal of Educational and Behavioral Statistics*, 45(1):5–31.
- Khajah, M. M., Huang, Y., González-Brenes, J. P., Mozer, M. C., & Brusilovsky, P. (2014). Integrating knowledge tracing and item response theory: A tale of two frameworks. In *CEUR Workshop Proceedings*, volume 1181, pages 7–15. University of Pittsburgh.
- Krivich, E., Hooshyar, D., Šír, G., Yang, Y., Bauters, M., Hämäläinen, R., & Kärkkäinen, T. (2025). A systematic review of deep knowledge tracing (2015–2025): Toward responsible AI for education. Preprints.
- Lee, W.-C. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement*, 47(1):1–17.
- Lee, W.-C. & Harris, D. J. (2025). Reliability in educational measurement. In Cook, L. L. & Pitoniak, M. J., editors, *Educational Measurement*, pages 277–381. Oxford University Press, 5 edition.
- Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., & Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.
- R Core Team (2026). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, 14(2):249–255.
- Rudner, L. M. (2000). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation*, 7(1):14.
- Shimada, D. & Okada, K. (2025). Reliability coefficient for Bayesian knowledge tracing models. *Technology, Knowledge and Learning*.
- Templin, J. & Bradshaw, L. (2013). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, 78(2):251–268.
- Templin, J. & Henson, R. (2009). Quantifying reliability in diagnostic classification models. In *Annual Meeting of the National Council on Measurement in Education*, San Diego, CA.
- Thompson, W. J., Clark, A. K., & Nash, B. (2019). Measuring the reliability of diagnostic mastery classifications at multiple levels of reporting. *Applied Measurement in Education*, 32(4):298–309.

- Thompson, W. J., Nash, B., Clark, A. K., & Hoover, J. C. (2023). Using simulated retests to estimate the reliability of diagnostic assessment systems. *Journal of Educational Measurement*, 60(3):455–475.
- Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement*, 52(4):457–476.
- Yamaguchi, K. & Templin, J. (2022). Observed score reliability indices in diagnostic classification models. *Behaviormetrika*, 49(1):47–68.
- Yudelson, M. V. (2017). HMM-scalable. Computer software. GitHub.
- Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013). Individualized Bayesian knowledge tracing models. In *Artificial Intelligence in Education*, pages 171–180. Springer.