

Fractional Information Gain: A Personalization Metric for Black-Box Predictive Models

S. Thomas Christie

Renaissance Philanthropy
thomas.christie@renphil.org

Markus Hauru

Renaissance Philanthropy
markus.hauru@renphil.org

Anna N. Rafferty

Carleton College
arafferty@carleton.edu

Abstract

Black-box knowledge tracing models are commonly deployed to drive personalization in on-line learning platforms and are typically evaluated using classification metrics such as AUC, accuracy, and F1 score. However, a model that predicts item responses using only each item's proportion correct in the training set achieves AUC up to 0.72 and accuracy up to 0.84 on widely used benchmark datasets, despite using no information about individual students' response histories. Inspired by the concept of marginal reliability in psychometrics, we introduce *Fractional Information Gain* (FIG), an information-theoretic evaluation metric for black-box predictive models of student item responses. FIG measures the fraction of a student's response uncertainty resolved by a trained model relative to the item-only baseline. FIG equals 0 when the model adds no information beyond item base rates, and 1 when held-out responses are perfectly predicted. FIG is applicable to any model that outputs probabilities and is sensitive to calibration errors. We characterize FIG on synthetic and real data, compare it to AUC for both the item-only baseline and trained models on four benchmark datasets, and describe the operational affordances that FIG inherits from its reliability-like construction. FIG imports the conceptual benefits of score reliability into the prediction-oriented framework of ML-driven educational systems.

1 Introduction

Personalizing a student's learning experience requires differentiating student interventions: students exhibiting meaningfully different behavioral evidence of learning should receive different interventions. While predictive models are often used to support personalization, the dominant evaluation metrics used to assess model quality (accuracy, AUC, and F1 score) were developed for general-purpose classification and do not clearly indicate a model's capacity to differentiate among students.

In both real and simulated datasets, models using only item correctness rates in the training dataset (e.g., 65% of responses to item j were correct) produce predictions (e.g., $p_j = 0.65$) much better than random (Figure 2), despite reflecting no information about any individual student. We refer to this predictor as the **item-only** baseline and use it throughout as a no-personalization reference point. Computing metrics for this baseline highlights that above-random values for reported metrics are not, on their own, evidence that a model has learned anything about individual students.

Questions of measurement quality have traditionally been decomposed into *validity* (is the right thing being measured?) and *reliability* (how much error is in the measurement?). Since classification metrics can produce high scores in the absence of individual student response histories, they are not by themselves valid measures of a model's ability to support personalization. For a valid metric, the baseline score would signal that the model has no ability to personalize, while the maximum score would indicate perfect knowledge of the student's state. This mirrors the family of reliability-related constructs in psychometrics (test-retest, parallel-forms, marginal reliability in item response theory) that are operationalized differently but share a common $[0, 1]$ scale, where 0 indicates "no information about the latent attribute" (Furr, 2021) and values near 1 indicate that scores are nearly free of noise. Whereas reliability > 0 signals that scores reflect student ability, AUC > 0.5 or above-chance accuracy does not have the same implication, sensitive as they are to item base rates.

Drawing inspiration from psychometrics, we propose *Fractional Information Gain* (FIG), an information-theoretic metric scaled like reliability but computed in the predictive validity space familiar to machine learning practitioners. For a student's observed responses X and target responses Y ,

$$\text{FIG} = \frac{I(X; Y)}{H(Y)} = 1 - \frac{H(Y | X)}{H(Y)}, \quad (1)$$

where H denotes entropy and $H(Y)$ is computed under training-set item base rates. FIG quantifies the proportion of uncertainty about responses Y resolved by knowledge of a student’s history X : 0 with no per-student uncertainty reduction and 1 at perfect prediction. FIG provides an interpretable, student-level, scaled signal for black-box predictive models analogous to the signal psychometric reliability provides for latent variable models.

Contributions. We (i) introduce FIG, a student-level personalization metric that removes the item-level base rates inflating AUC and other classification metrics; (ii) demonstrate the baseline inflation in several metrics for commonly used datasets; (iii) characterize the sensitivity of FIG’s empirical estimator, FIG-V, to model bias and miscalibration; (iv) show that FIG-V behaves as expected in a model trained on simulated data, localizing personalization gain to skill-relevant inputs; and (v) describe operational affordances provided by FIG’s reliability-like construction. We release FIG as the Python package `fractional-information-gain` on [PyPI](#) and [GitHub](#).

2 Background

2.1 Predictive validity metrics

Binary student-response models predict $P(Y_i = 1 | X)$ for held-out responses Y_i using a student’s response history X . Common model quality metrics include accuracy and F1 score (threshold-dependent), AUC (rank-based, threshold-free), Brier score and RMSE (sensitive to probability calibration), and binary cross-entropy (BCE, the negative Bernoulli log-likelihood used as a training loss). Several recent papers argue against overreliance on individual metrics in educational contexts, citing issues including base-rate effects and insensitivity to model calibration ([Baker et al., 2024, 2025](#); [Kwegyir-Aggrey et al., 2023](#)). AUC summarizes prediction ordering while F1 score and accuracy depend on thresholds; all three are therefore narrow and occasionally pathological evaluation metrics for models outputting probabilities ([Pelánek, 2015](#); [Kwegyir-Aggrey et al., 2023](#)). Brier score, RMSE, and BCE are sensitive to model calibration, but do not have a baseline-anchored scale and vary across datasets, and so are infrequently reported.

Our concerns are threefold: (i) all commonly used metrics improve with knowledge of item base rates alone, sometimes substantially; (ii) none has a fixed reference point that distinguishes “the model can make inferences using a student’s history” from “the model has learned about items”; (iii) as typically reported, metrics pool across all responses in the test set, a practice that can bias metrics toward performance for students with many responses.

2.2 Reliability in psychometrics

In psychometrics, analogous challenges are addressed by the notion of *reliability*. In item response theory (IRT) and related latent variable models, an estimate of a student-level parameter such as ability is referred to as a *score*. Score reliability summarizes the noise present in score estimates. Operationally, reliability is a family of measurement-related constructs rather than a single quantity ([Furr, 2021](#)).

We highlight two examples to make the relationship between uncertainty and reliability explicit. In classical test theory ([DeVellis, 2006](#)), an observed score¹ $O = T + E$ decomposes into the true score T and an error E , with reliability given by:

$$\rho_{OO'} = \frac{\text{Var}(T)}{\text{Var}(O)} = 1 - \frac{\text{Var}(E)}{\text{Var}(O)}. \quad (2)$$

Item response theory ([Cai et al., 2016](#)) expresses the idea via marginal reliability,

$$\begin{aligned} \rho_\theta &= 1 - \frac{\mathbb{E}_X[\text{Var}(\theta | X)]}{\text{Var}(\theta)} \\ &\approx \frac{1}{S} \sum_{s=1}^S \left[1 - \frac{\text{Var}(\theta | X^{(s)})}{\text{Var}(\theta)} \right] \end{aligned} \quad (3)$$

using the model’s posterior uncertainty $\text{Var}(\theta | X)$ over student parameter θ conditioned on a student’s responses X , where the expectation is estimated by averaging over S observed students with responses $X^{(1)}, \dots, X^{(S)}$. All formulations encode reliability on a fixed $[0, 1]$ scale as

$$\text{Reliability} = 1 - \frac{\text{Residual uncertainty}}{\text{Total uncertainty}}.$$

Reliability quantifies the precision with which student-level parameters can be inferred, but it cannot be used with black-box predictive models, which lack interpretable student-level parameters.

¹We write O for the observed score rather than the conventional X , which we reserve for a student’s response history.

3 Fractional Information Gain

Fractional Information Gain (FIG) is an information-theoretic measure of how much a trained model’s predictions resolve uncertainty about a student’s responses relative to an item-only baseline. It applies to any probability-producing item response model, sits on a fixed scale anchored at zero (no personalization) and one (perfect predictions), and parallels the structural forms of psychometric reliability (see Section 2.2).

FIG is built on Shannon entropy (Shannon, 1948). For a discrete random variable X with probability mass function P , the entropy $H(X) = -\sum_i P(X=i) \log P(X=i)$ quantifies uncertainty; the conditional entropy $H(Y|X)$ measures the remaining uncertainty in Y after observing X ; and the mutual information $I(X;Y) = H(Y) - H(Y|X)$ quantifies the reduction in uncertainty in Y due to observing X . FIG is the fraction of a student’s response uncertainty that their history X resolves.

Operationally, let $Y = (Y_1, \dots, Y_n)$ be a set of unobserved binary responses for a single student, let q_i be the item associated with response Y_i , and let X be the student’s observed behavior. Let $p_{Di} = \frac{\# \text{ correct responses to } q_i}{\# \text{ total responses to } q_i}$ be the base rate of correct responses to item q_i in a training set D , and $p_{Mi} = P(Y_i = 1 | X)$ be the trained model M ’s prediction. Given only item statistics, the modeler’s prior uncertainty about a given student’s responses is

$$H_D(Y) = \sum_{i=1}^n h(p_{Di}), \quad (4)$$

where $h(p) = -p \log p - (1-p) \log(1-p)$ is the binary entropy function. While the general form of FIG (Equation 1) can be computed using the model’s conditional entropy, to anchor the metric in held-out observations $Y_i = y_i$ we instead compute the binary cross-entropy of the model’s predictions against those observations,

$$\hat{H}_M(Y|X) = \sum_{i=1}^n \text{BCE}(y_i, p_{Mi}), \quad (5)$$

where $\text{BCE}(y, p) = -y \log p - (1-y) \log(1-p)$.

We then define:

$$\text{FIG-V} = 1 - \frac{\hat{H}_M(Y|X)}{H_D(Y)}. \quad (6)$$

FIG-V (V for validation) is an empirical estimator of FIG that uses binary cross-entropy on

held-out responses and is suitable for predictive validation during and after model training. In expectation, FIG-V is a lower bound on FIG that is tight when the model’s predictions match the true response probabilities.

Pooled vs. per-student FIG-V. Equation 6 defines FIG-V for a single student. To summarize across a population of S students, where each student s has held-out responses $Y^{(s)}$ and observed behavior $X^{(s)}$, we consider two natural aggregations. *Pooled FIG-V* sums numerator and denominator across students before forming the ratio, $1 - \sum_s \hat{H}_M(Y^{(s)} | X^{(s)}) / \sum_s H_D(Y^{(s)})$; it is response-weighted. *Per-student FIG-V* computes FIG-V separately for each student and averages with equal weight per student, analogous to Equation 3:

$$\text{FIG-V}_{\text{per-student}} = \frac{1}{S} \sum_{s=1}^S \left[1 - \frac{\hat{H}_M(Y^{(s)} | X^{(s)})}{H_D(Y^{(s)})} \right]. \quad (7)$$

The two forms diverge when students answer widely varying numbers of items, as is typical in real knowledge tracing (KT) datasets. We argue that per-student FIG-V is the more appropriate summary when the goal is to evaluate differential decision support for individual students.

Interpretation. In expectation, FIG-V equals zero when the model’s negative log-likelihood on held-out responses matches that of the item-only baseline ($\hat{H}_M(Y|X) \approx H_D(Y)$); equals 1 when held-out responses are perfectly predicted ($\hat{H}_M(Y|X) \approx 0$); and is negative when the model assigns lower likelihoods to observed responses than item base rates do. FIG itself (Equation 1) is nonnegative because mutual information is nonnegative. A negative FIG-V value for a trained model is diagnostic of systematic miscalibration, overfitting, or a poor training procedure.

Equations 1 and 6 are structurally similar to reliability Equations 2 and 3: each expresses model quality as one minus the ratio of remaining-to-total uncertainty, where uncertainty is quantified as Shannon entropy in the former and score variance in the latter. FIG-V also resembles metrics such as the log-score skill score, predictive deviance reduction, and McFadden’s pseudo- R^2 (McFadden, 1974), defined as $R^2 = 1 - \frac{\text{NLL}_{\text{model}}}{\text{NLL}_{\text{baseline}}}$, where NLL is negative log-likelihood. By design, FIG-V uses per-item correctness entropy evaluated using the training set, explicitly quantifying personalization

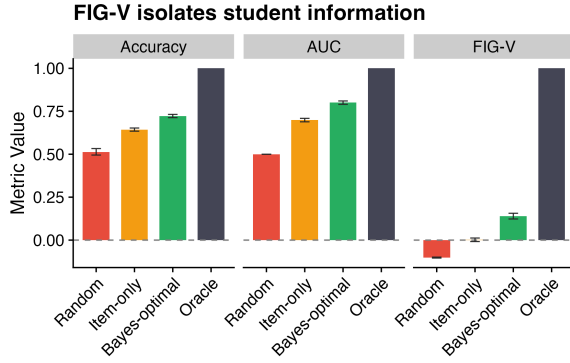


Figure 1: Per-student FIG-V is approximately 0 when item base rates are used as predictions, is negative for random predictions, and separates Bayes-optimal from oracle predictions far more sharply than the other metrics do. Error bars are 95% student-bootstrap CIs.

relative to a non-personalized baseline, rather than generic model improvement.

Selecting Y . Black-box predictive models do not directly produce skill estimates. Following the “market basket” tradition (Mislevy, 1998), in which a fixed, interpretable collection of items serves as the reporting scale, we suggest that Y be a meaningful collection of items related to a skill or construct of interest, for example, a standard in the Common Core or a more granular node in the Learning Commons knowledge graph (Learning Commons, 2025). When Y is the entire held-out set, FIG-V functions as a global model-quality summary analogous to AUC. When Y represents one of several skill-aligned item subsets, FIG-V can be used to produce multiple skill-specific signals without re-training.

4 Properties of FIG-V

We characterize FIG-V on IRT-simulated data, where ground-truth student parameters are available, and on four widely used KT datasets. Code to reproduce all our results is available at github.com/RenaissancePhilanthropy/FIG-AIME-2026.

4.1 Metric comparison using simulated data

We first compare FIG-V against accuracy and AUC using data simulated with a 1PL IRT model. While IRT models are latent variable models rather than black-box models, they are useful here because ground-truth response probabilities are known: $p(\text{correct} \mid \theta, d) = \sigma(\theta - d)$, with student abilities $\theta \sim \mathcal{N}(0, 1)$, item difficulties $d \sim \mathcal{N}(0, 1)$, where σ is the sigmoid function.

We created a training set of 5,000 students and a test set of 500 students, each responding to a random subset of 20 items from a 200-item pool. The training set is used to estimate baseline correctness rates for each item. We then make predictions for responses in the held-out set using four methods: (1) random ($p = 0.5$); (2) item-only ($p_i =$ proportion of training-set responses to item i that are correct); (3) Bayes-optimal ($p = \sigma(\theta - d)$ with known θ and d), representing the best predictions achievable by any model; and (4) an oracle that predicts every answer perfectly. As shown in Figure 1, predictions using only item statistics (orange bars) produce FIG-V ≈ 0 but better-than-random accuracy and AUC values. Compared to traditional metrics, FIG-V also highlights the distinction between Bayes-optimal predictions, which retain aleatoric uncertainty (green), and oracle predictions (dark gray).

4.2 Patterns persist with real data

The metric behavior seen in Figure 1 also appears in real data collected from students. Figure 2 shows accuracy, AUC, and FIG-V calculated for four standard KT datasets benchmarked in Liu et al. (2022) alongside two additional simulated datasets. The real datasets were split into training and testing sets, and item correctness rates were computed using the training set. ‘Item-only’ reflects predictions using the item base rates. ‘Trained SAKT’ shows metric values produced by Self-Attentive Knowledge Tracing (SAKT; Pandey and Karypis, 2019) models trained on each dataset. For the real datasets, ‘Reference’ gives the question-level performance of the best model in Liu et al. (2022) that uses only item identifiers as input (DKT+ for AS’15, AL’05, and BD’06; SAINT for NIPS34). As with Figure 1, ‘Bayes-optimal’ bars indicate performance when the data-generating probabilities are used to make predictions in simulated datasets.

Item-only AUC and accuracy are above random on every dataset, indicating that these metrics are not a pure measure of a model’s capacity to incorporate student histories into predictions. Critically, the datasets are ordered differently by AUC and by FIG-V: while a naive reading of AUC would indicate that models trained on AL’05 (Stamper et al., 2010a) and BD’06 (Stamper et al., 2010b) support the best personalization, FIG-V reveals that models trained on AS’15 (Heffernan and Heffernan, 2014; ASSISTments, 2015) and NIPS34 (Wang et al., 2020) resolve the largest fraction of uncer-

Prediction metrics for widely used datasets

Numbers give item-only values.

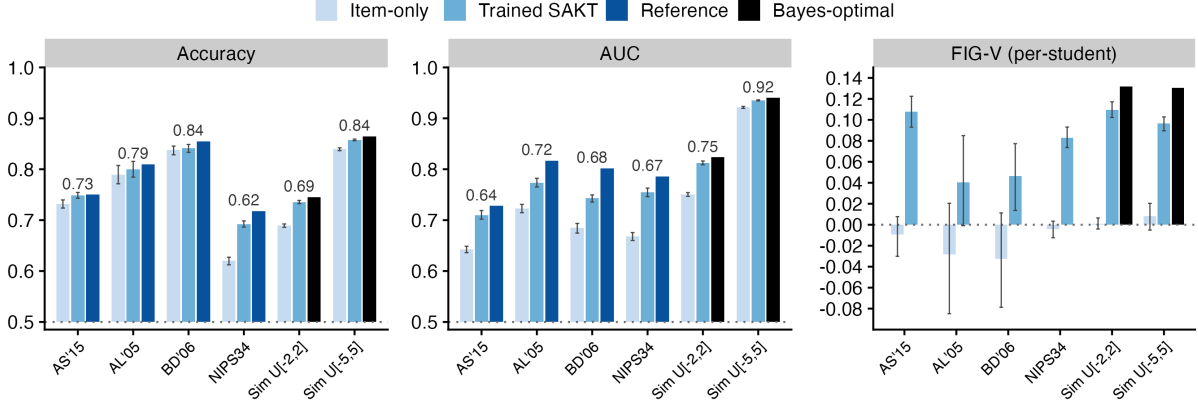


Figure 2: Metric values for predictions from the item-only baseline, a SAKT model trained on each dataset, and the best item-ID-only model reported in [Liu et al. \(2022\)](#). Accuracy and AUC are high even for the item-only baseline; FIG-V filters out the effect of item difficulty, showing that the trained model adds only a small margin. Error bars are 95% student-bootstrap CIs; the black bars are the Bayes-optimal ceiling on the simulated sets.

tainty about a student’s future responses from their history.

To illustrate the connection between properties of a dataset and resulting metric values, we simulated two additional datasets. Both are simulated from a 1PL IRT model with item difficulties $d \sim U[-2, 2]$ and $d \sim U[-5, 5]$, respectively, keeping $\theta \sim \mathcal{N}(0, 1)$. The large difference in baseline AUC (0.75 and 0.92) is attributable solely to item properties, while FIG-V remains approximately 0 for all datasets under the item-only condition. The Bayes-optimal predictions under the two simulated datasets show markedly different AUC values *despite both having perfect knowledge of student parameters*. In contrast, FIG-V values for these two datasets are similar.

4.3 Sensitivity to error and miscalibration

To investigate how different forms of model error affect each metric, we draw responses from a true IRT model and score deliberately misspecified predictions against them. Responses are generated once using $p(\text{correct} \mid \theta, d) = \frac{1}{1 + e^{-a(\theta - \nu - d)}}$ with discrimination parameter $a = 1$ and bias parameter $\nu = \theta - \hat{\theta} = 0$. We then generate misspecified predictions by modifying the bias ν to inject ability-estimation error and discrimination a to inject calibration error ($a < 1$ underconfident, $a > 1$ overconfident; $a = 1, \nu = 0$ recovers the true model), recomputing every metric at each setting. The true student θ is fixed at 0, item difficulties are drawn as $d \sim \mathcal{N}(0, 1)$ for 100 items, and metrics are averaged over 5,000 simulated response vectors

for each configuration.

Results in Figure 3 show FIG-V’s relatively strong sensitivity to both model miscalibration and bias. In contrast, AUC is constant across variations in both miscalibration and bias, emphasizing its function as a summary of rank ordering. Accuracy and F1 vary with bias but are insensitive to a , since scaling by a does not change which predictions exceed 0.5. FIG-V is maximized for well-calibrated models (here, $a = 1$) with zero student error ($\theta - \hat{\theta} = 0$), and decreases monotonically when either parameter drifts. Unlike RMSE, FIG-V’s anchored scale clearly indicates when the model reduces uncertainty relative to the item-only baseline (FIG-V > 0) and when it does not (FIG-V < 0), and is interpretable across students and datasets.

4.4 Skill-specific predictions

As mentioned in Section 3, intentionally selecting Y is a practical affordance allowing the modeler to quantify predictive capacity for different subsets of items using a single model. To illustrate this, we trained a SAKT model on a five-skill synthetic dataset (training set: 200-item pool with 40 items per skill, 5,000 students answering all items, skill-specific abilities $\sim \mathcal{N}(0, 1)$, item difficulties $\sim U[-2, 2]$; skills independent by construction). Item ordering is randomized and student skills remain static. In each N -item sequence, items in positions $1 \dots k$ are used as input to predict items $k + 1 \dots N$.

Figure 4 illustrates changes in FIG-V as the

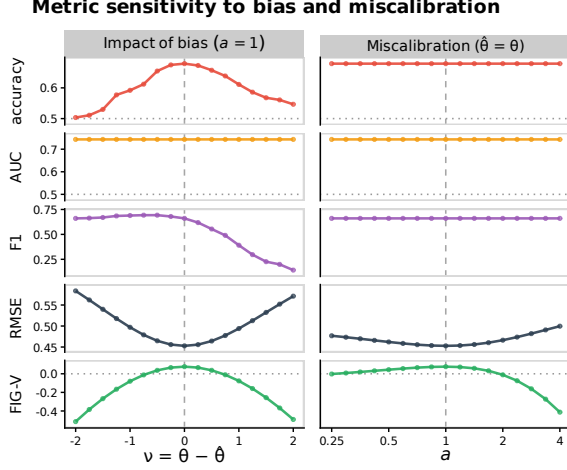


Figure 3: Behavior of five evaluation metrics on systematically incorrect predictions. Metrics compare responses generated from $p(\text{correct} \mid \theta = 0, d) = \frac{1}{1 + e^{-a(\theta - \nu - d)}}$ with $a = 1$ and $\nu = 0$ against predictions produced with $a \in [0.25, 4]$ and $\nu \in [-2, 2]$. (left) AUC is constant under biased predictions; FIG-V, RMSE, and accuracy are optimized at zero bias, whereas F1 is highest for negative ν , since F1 ignores true negatives and so rewards over-predicting correct responses. (right) F1, accuracy, and AUC are insensitive to miscalibration, while RMSE and FIG-V are optimized for calibrated predictions.

model is given k of a held-out student’s responses to an arbitrarily chosen skill (Skill 3) as its input X , highlighting two expected properties. First, a greater number of Skill 3 observations produces stronger predictions on held-out Skill 3 responses. Second, the gain is restricted to items aligned with the observed skill. These properties are not exclusive to FIG-V, as other metrics would vary in similar ways. However, the interpretation is distinctive: FIG-V quantifies the proportion of available uncertainty about the items in Y resolved by the observations in X .

5 Discussion

When using predictive models to drive student personalization, the salient question is “what fraction of uncertainty about a student’s future responses Y is reduced by observed responses X ?” As an answer, we introduce a principled, operational quantification called Fractional Information Gain, and an estimator, FIG-V, for model validation. FIG and FIG-V share the $[0, 1]$ scale of psychometric reliability, with an analogous interpretation, though they are expressed in proportional entropy rather

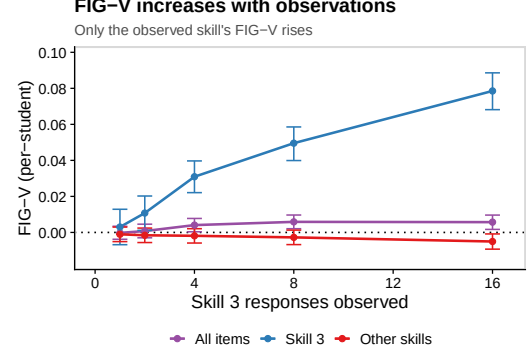


Figure 4: Per-student FIG-V increases as more responses are observed for a target skill, but only for that skill. On a five-skill synthetic dataset (independent abilities), a SAKT model is shown k of a held-out student’s Skill 3 responses and predicts the rest. FIG-V on held-out Skill 3 items climbs with k , while FIG-V on the *other* (independent) skills is flat or slightly declining near zero. Observing one skill carries no information about the others. “All items” (FIG-V over the full held-out set) rises only modestly, as it is dominated by the other-skill items. Points are per-student FIG-V; error bars are 95% student-bootstrap CIs.

than proportional variance.

Change in baseline comparison. The item-only results in Figures 1 and 2 highlight that traditional predictive validity metrics implicitly compare against a misleading reference point. There are two natural uninformative predictors when modeling student responses: a student responding at chance level ($p = 0.5$), and predictions using item base rates alone (which arise, for instance, under posterior collapse in variational autoencoders; Christie et al., 2023; Lucas et al., 2019). We argue that the second is the more meaningful anchor in the context of educational personalization, as models using item statistics to predict correctness can otherwise appear to make strong predictions without any incorporation of student response histories, with AUCs of 0.64–0.72 in widely used datasets and higher in common simulation paradigms. Practitioners *could* compute AUC and accuracy for predictions using item base rates, but this is not common practice, and baseline metrics would vary by dataset. In contrast, FIG-V includes this computation by construction, with zero indicating “no gain from personalization” regardless of the dataset.

An information-theoretic metric. The question “how much does a student’s behavior trajectory tell us about future item responses?” is fundamentally a question about uncertainty reduction, and mutual

information is its principled quantification. Scaling by baseline entropy gives FIG-V an interpretable $[0, 1]$ scale analogous to reliability, with the answer given in appropriate units (proportion of uncertainty resolved). By contrast, AUC and accuracy summarize quantities (pairwise ordering, threshold-based correctness) that are related to, but not clear signals of, a model’s capacity to personalize; both can be high for predictions that incorporate no student information.

Future work: a reference-free variant. Like traditional model validation metrics, FIG-V is computed by comparing predictions to held-out observations. In principle, the information-theoretic formulation in Equation 1 could also be applied to estimate uncertainty reduction for an individual student without held-out data. As a psychometric analogue, the IRT standard error of measurement (SEM) is a student-level quantity summarizing the modeler’s confidence in an individual score; the Bayesian analogue is the standard deviation of the score posterior. In similar fashion, rather than estimating $H(Y | X)$ with $\hat{H}_M(Y | X)$ on held-out data, one can compute the model’s own predictive entropy directly. Using this in Equation 1 rather than BCE gives a reference-free variant we call *FIG-C* (C for confidence), which plays a role analogous to SEM: it summarizes the modeler’s confidence about an individual student’s future responses, computable from the model’s beliefs alone without observed held-out behavior. Provided the model is well calibrated, FIG-C and FIG-V agree in expectation. Potential applications include adaptive item selection, using expected FIG-C gain as an analogue of Fisher information, and information-theoretic stopping rules for assessment based on FIG-C thresholds. A full treatment, including calibration requirements, item-selection validation, and threshold design for adaptive stopping, is left to future work.

Limitations

Point estimates for item base rates. FIG-V normalizes the model’s held-out cross-entropy by an entropy quantity derived from item base rates (Equation 4). Item base rates are sample statistics whose precision depends on sample size. Consequently, FIG-V estimates can be noisy if they involve items with few responses in the training set. Practitioners may either omit items with small counts from analysis (a common psychometric practice) or apply Beta-prior shrinkage when com-

puting proportions, which adds a pseudo-correct and b pseudo-incorrect observations to each item. The prior’s influence diminishes as responses accumulate. The latter method is implemented in the software package² and accommodates user-provided a and b . Items with little data could also have priors informed by similar items (say, items of similar type or alignment), but we leave this as future work. In each case, $H_D(Y)$ should represent the modeler’s belief about a student’s likely responses before they are observed.

Aleatoric uncertainty. FIG-V’s $[0, 1]$ scale spans the range from adding no information about held-out responses to predicting them perfectly. For an observed FIG-V below 1, it is unclear how much of the remaining gap could be reduced by a better model or more data. In other words, the metric does not, and cannot, distinguish between epistemic uncertainty that can be reduced with more data and a better model, and aleatoric uncertainty that is irreducible noise inherent in the data-generating process. This is a challenge for every post hoc predictive validity metric, as the distinction can only be made with a forward model of data generation: for example, the Bayes-optimal predictions for simulated data in Figure 2 have zero epistemic uncertainty and mark the ceiling of achievable performance. For this reason, we encourage practitioners to simulate student data from realistic generative models in order to explore the conditions under which this ceiling can be approached.

Acknowledgments

The authors thank Josh Mitton and Prarthana Bhattacharyya at Eedi and Zachary Levonian at Renaissance Philanthropy for valuable feedback. The authors also thank the Learning Engineering Virtual Institute, a program of Renaissance Philanthropy, for supporting this work.

References

- ASSISTments. 2015. 2015 ASSISTments skill builder data. <https://sites.google.com/site/assistmentsdata/datasets/2015-assistments-skill-builder-data>.
- Ryan Baker, Nigel Bosch, Stephen Hutt, Andres Zambrano, and Alex Bowers. 2024. On fixing the right problems in predictive analytics: AUC is not the problem. *arXiv preprint arXiv:2404.06989*.

²github.com/RenaissancePhilanthropy/fractional-information-gain

- Ryan Baker, Caitlin Mills, and Jaeyoon Choi. 2025. The difficulty of achieving high precision with low base rates for high-stakes intervention. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 790–796.
- Li Cai, Kilchan Choi, Mark Hansen, and Lauren Harrell. 2016. Item response theory. *Annual Review of Statistics and Its Application*, 3(1):297–321.
- S. Thomas Christie, Hayden Johnson, Carson Cook, Garron Gianopulos, and Anna N. Rafferty. 2023. LENS: Predictive diagnostics for flexible and efficient assessments. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*, pages 14–24.
- Robert F. DeVellis. 2006. Classical test theory. *Medical Care*, 44(11):S50–S59.
- R. Michael Furr. 2021. *Psychometrics: An introduction*. SAGE Publications.
- Neil T. Heffernan and Cristina Lindquist Heffernan. 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.
- Kweku Kwegyir-Aggrey, Marissa Gerchick, Malika Mohan, Aaron Horowitz, and Suresh Venkatasubramanian. 2023. The misuse of AUC: What high impact risk assessment gets wrong. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’23, pages 1570–1583, New York, NY, USA. Association for Computing Machinery.
- Learning Commons. 2025. Learning Commons Knowledge Graph. <https://github.com/learning-commons-org/knowledge-graph>. GitHub repository. Accessed 2026-02-08.
- Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, Jiliang Tang, and Weiqi Luo. 2022. pyKT: A Python library to benchmark deep learning based knowledge tracing models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- James Lucas, George Tucker, Roger B. Grosse, and Mohammad Norouzi. 2019. Don’t blame the ELBO! A linear VAE perspective on posterior collapse. *Advances in Neural Information Processing Systems*, 32.
- Daniel McFadden. 1974. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, New York.
- Robert J. Mislevy. 1998. Implications of market-basket reporting for achievement-level setting. *Applied Measurement in Education*, 11(1):49–63.
- Shalini Pandey and George Karypis. 2019. A self-attentive model for knowledge tracing. In *12th International Conference on Educational Data Mining, EDM 2019*, pages 384–389. International Educational Data Mining Society.
- Radek Pelánek. 2015. Metrics for evaluation of student models. *Journal of Educational Data Mining*, 7(2):1–19.
- Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.
- J. Stamper, A. Niculescu-Mizil, S. Ritter, G. J. Gordon, and K. R. Koedinger. 2010a. Algebra I 2005–2006. Challenge data set from KDD cup 2010 educational data mining challenge. <http://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp>.
- J. Stamper, A. Niculescu-Mizil, S. Ritter, G. J. Gordon, and K. R. Koedinger. 2010b. Bridge to algebra 2006–2007. Challenge data set from KDD cup 2010 educational data mining challenge. <http://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp>.
- Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Yordan Zaykov, José Miguel Hernández-Lobato, Richard E. Turner, Richard G. Baraniuk, Craig Barton, Simon Peyton Jones, Simon Woodhead, and Cheng Zhang. 2020. Diagnostic questions: The NeurIPS 2020 education challenge. *arXiv preprint arXiv:2007.12061*.