A Problem that Shouldn't be Skipped: Problem Skipping Limits the Accuracy of Ability Estimates in Online Learning

Abstract

Large-scale personalized learning systems rely on accurate and fair student ability estimations, yet common unidimensional models often neglect meta-cognitive processes. As such, they can provide biased estimates that put a limit on accuracy and fairness. This study introduces a multidimensional approach using the IRTree framework, which models problem-skipping as a distinct latent trait alongside accuracy. Problem-skipping is a common meta-cognitive manifestation that can reflect self-regulation and strategic decision-making. Our results demonstrate that problem-skipping reflects meaningful individual differences, with patterns of problem-skipping varying in their impact on performance. Notably, while frequent problem-skipping was associated with lower accuracy, it also revealed adaptive strategies for some learners. This suggests that problem-skipping is signal, not noise. To ensure that learning analytic tools are supported by accurate and fair measurement models, we suggest several ways to account for problem-skipping when estimating student ability.

Keywords adaptive learning, learner performance modeling, problem-skipping, item response tree, K-12 education, fairness

1 Introduction

For the sake of better understanding and improving students' processes during online learning, educational data mining relies on measurement models that accurately and fairly assess student ability. Measurement models underlying large-scale personalized learning systems are primarily unidimensional, *i.e.*, the primary observations driving the latent trait estimation are whether or not a response was correct (here referred to as accuracy). However, many platforms offer response options beyond simply giving the answer, such as requesting hints or skipping questions (here referred to as problem-skipping). While such behaviors have been used in the past to predict performance, the latent traits underlying problem-skipping remain unclear. This study explores whether skipped responses represent a distinct cognitive process, separate from accuracy, in estimating learner ability.

Deciding whether to skip a problem or not requires the student to self-assess their likelihood of success and allocate their effort accordingly (Winne, 2017). This meta-cognitive process, consisting of monitoring and regulation, is crucial for successful learning (Colantonio et al., 2024; Fleur et al., 2021; Rabinowitz, 2017), particularly in online settings (Fidan & Koçak Usluel, 2024; Kim et al., 2009; Shields et al., 2024; Zhao & Ye, 2020). The extent to which meta-cognitive tools are used differs across learners, as a function of individual differences in, for example, prior knowledge (Taub et al., 2014), motivation (Karlen, 2016; Sungur, 2007), and anxiety (Matthews et al., 1999; Wells, 1995). Differential patterns in problem-skipping may also be induced by avoidance strategies stemming from reduced effort or affective states, such as anxiety and boredom, which influence engagement and self-regulation through their impact on motivation and perceived control (Pekrun & Perry, 2013). Compounding this issue, students do not always know that they need help (Aleven & Koedinger, 2000). If individual differences in the use of meta-cognitive strategies are ignored, systemic patterns of problem-skipping behavior underlying such traits could lead to confounds in ability estimates. By explicitly modeling problem-skipping behavior, we capture an important dimension of self-regulated learning that traditional unidimensional models overlook.

Research on problem-skipping in online learning is sparse, and existing studies reveal a mixed portrait of its relationship with accuracy. On the one hand, results from a large-scale intervention study demonstrated that limiting access to a problem-skipping option increased effortful practice and improved learning in an online mathematics learning platform (Savi et al., 2018, 2023). Hint-requests have also been found to correlate negatively with pre- and post- mathematics test scores, and positively with mathematical errors during online learning (Norum et al., 2024). These studies point to a negative relationship between problem-skipping and accuracy. On the other hand, in the same study (Norum et al., 2024), the authors found evidence for individual differences in the extent to which the hint-request option was used, and highlighted that a proportion of students which were identified to need the most help relied the least on help-seeking strategies. Similarly, an eye-tracking study found that students with lower accuracy paid less attention to hints (Conati et al., 2013). Further studies have found evidence for individual differences in hint-processing but show that these individual differences are not related to learning gains (Goldin et al., 2012, 2013). Further, studies aiming to classify online learning behavior as "off-task" or disengaged (Baker, Corbett, & Koedinger, 2004; Baker, Corbett, Koedinger, & Wagner, 2004; Baker et al., 2005; Beck, 2004) reveal a group of students who misuse response options such as skipping. These findings suggest that problem-skipping behaviors may not be explained by student ability alone.

1.1 Unidimensional measurement models

Problem-skipping has previously been incorporated into ability assessment, largely within Bayesian Knowledge Tracing (BKT) frameworks. Since its introduction (Corbett & Anderson, 1995), where problem-skipping behavior is treated the same as an incorrect response, there have been several extensions which treat problem-skipping differently. For example, the partial-credit model (Ostrow et al., 2015; Wang & Heffernan, 2013; Wang et al., 2010), gives some, but not full, credit for hint-seeking or multiple attempts to a problem. Other implementations include incorporating the history of help-seeking to improve the prediction accuracy of guesses and slips (Baker et al., 2008) or detecting affective states (Corrigan et al., 2015). A comprehensive overview of such proposed

enhancements to the classical BKT model is provided by Šarić-Grgić et al. (2024).

Another commonly used measurement model for assessing student ability is Item Response Theory (IRT). In this model, student ability is assessed based on the likelihood of a correct response, given specific properties of the current test item, and the student's current ability. In this paper, we use extensions of the 1-Parameter Logistic Model (1PL; also referred to as the Rasch Model (Rasch, 1993)) to account for hierarchical decision processes. In the 1PL model, the probability that a student with ability θ_p answers item i correctly is given by the function:

$$P(X_i = 1 \mid \theta_p, \beta_i) = \frac{1}{1 + e^{-(\theta_p - \beta_i)}} \tag{1}$$

where β_i is the difficulty parameter for item i, which represents at what ability level the student has a 50% chance of responding to the item correctly. In BKT terms, the θ parameter (student ability) represents the probability of moving from an unlearned to a learned state, and the β parameter (item difficulty) represents an item-specific probability of moving from a learned state back into an unlearned state. This bridge between IRT and BKT models of student ability has been formalized by (Deonovic et al., 2018). Although the current paper deals with data obtained under an IRT measurement model, results will be relevant for online learning systems utilizing a BKT-framework.

Importantly, IRT and BKT models share the feature that ability is measured on a unidimensional scale. Both measurement models thus assume that problem-skipping is driven by the same latent trait as the accuracy of a response. Previous work has demonstrated that ignoring missing responses in IRT approaches can lead to bias on person- and item-parameter estimates (Debeer et al., 2017; Finch, 2008; Holman & Glas, 2005; Mislevy & Wu, 1996; Pohl et al., 2014). Similarly, prediction accuracy was found to increase in BKT models when disengaged responses (including long sequences of skipped problems) were excluded from the ability estimation (Gorgun & Bulut, 2022). These findings support the notion that problem-skipping might be indicative of a separate meta-cognitive state which is not directly related to ability, and signal a need to test the assumption

that problem-skipping and accuracy can be measured on the same dimension.

1.2 A multidimensional approach

Failure to account for individual differences in response patterns is problematic because it may bias the validity of the trait measurement and other constructs to be estimated, particularly when the response style is related to the trait to be measured. Therefore, a range of research is dedicated to reliably identifying and controlling for response styles to produce ability estimates which are as unbiased as possible. One established way is to model the response process as a hierarchical decision tree, referred to as an IRTree model (De Boeck & Partchev, 2012), where the content and the response format of the item are both assumed to contribute to the observed response (Böckenholt, 2012; De Boeck & Partchev, 2012; Thissen-Roe & Thissen, 2013).

IRTree models can be used to model skipped responses in an educational context. In specific, the model assumes a two-stage process where, first, the item is responded to or not, and second, the item is responded to correctly or incorrectly. This process can be visualized as a tree-branching model with two nodes (Figure 1), and the response of person p on item i can be coded as two binary variables, $Y^{(1)}$ and $Y^{(2)}$, where

$$Y^{(1)} = \begin{cases} 1 & \text{if the person attempts item } i \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$Y^{(2)} = \begin{cases} 1 & \text{if the person solves item } i \text{ correctly,} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The tree model is a multidimensional IRT model, where the probability of the response of person p on item i being in category m can be formulated as:

$$\pi(x_{pi} = m \mid \theta_p, \beta_i) = \prod_{n=1}^{N} \left[\left(\frac{1}{1 + e^{-(\theta_p^{(n)} - \beta_i^{(n)})}} \right)^{t_{mn}} \right]^{q_{mn}} \tag{4}$$

Each node in the tree is modeled with a 1PL model. Depending on where you are in the tree, the probability is computed as the product across all internal nodes involved in the respondent's path through the tree, leading to a specific response category. The terms t_{mn} and q_{mn} decide whether a node will contribute to the probability of category m, in our case question-mark, incorrect or correct. t_{mn} determines whether an internal node n is relevant for the probability calculation of response category m. It equals 1 if the node is relevant to category m (*i.e.*, contributes to the response), and 0 otherwise. q_{mn} determines whether the node n is active for the probability calculation of category m). When $q_{mn} = 1$, the node will actively contribute to the probability of category m, and when $q_{mn} = 0$, the node will be ignored for that category.

For example, consider a student who does *not* skip item i ($Y^{(1)} = 1$), and then responds correctly ($Y^{(2)} = 1$): at node 1, their probability of attempting the item (not skipping) is modeled as a 1PL model (equation 1), and at node 2, their probability of correctly answering item i follows a second 1PL model (equation 1). The final response depends on successfully moving through both nodes, so the overall probability of providing a correct answer, $P(Y^{(1)} = 1, Y^{(2)} = 1)$, is the product of these two probabilities. At both nodes, t_{mn} and q_{mn} are both equal to 1, as they are relevant and active for the final response category. Inversely, for a student who *does* skip the problem (*i.e.*, $Y^{(1)} = 0$), node 1 is relevant ($t_{m1} = 1; q_{m1} = 1$) but node 2 is not ($t_{m2} = 0; q_{m2} = 0$). Thus, the overall response probability is only the probability of skipping ($P(Y^{(1)} = 0)$), modeled by a 1PL model, and they do not proceed further in the tree.

The main advantage with this modeling approach is that the unidimensionality of the decision process can be formally tested (De Boeck & Partchev, 2012). The two-node branching model depicted in Figure 1 gives rise to two latent traits on the person and item side respectively:

1. node 1: θ_p^1 (individual propensity to skip an item) and β_i^1 (propensity for item to be skipped), and

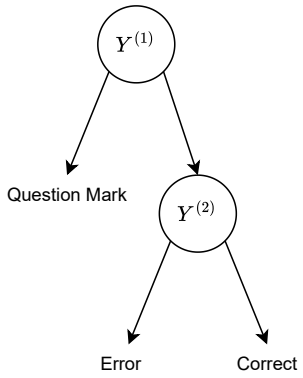

Figure 1: A binary decision tree (IRTree model) with 2 nodes reflecting the decision to respond or not (Y_1) and whether the response is correct or not (Y_2).

2. node 2: θ_p^2 (individual ability) and β_i^2 (item difficulty).

Now, the correlation between θ^1 and θ^2, and β^1 and β^2 can be calculated. That is, how much do skipped responses depend on respondents or items, respectively? The unidimensional model assumes that these correlations are negligible, *i.e.*, for individuals, problem-skipping and accuracy are the same, and the likelihood for an item to be skipped is equal to its estimated difficulty. This assumption is tested by comparing the unidimensional to the multidimensional (IRTree) model.

Previous work utilizing IRTree models has found support for problem-skipping and accuracy as distinct processes in educational assessment data obtained from traditional pen-and-paper tests (Debeer et al., 2017; Holman & Glas, 2005; Huang, 2020; Okumura, 2014; Pohl et al., 2014). A simulation study found that a 1PL model results in bias in person- and item-parameters when data is missing not at random, and that modeling the missingness mechanism with an IRTree model significantly reduces this bias (Debeer et al., 2017). This was especially the case for students who skip many responses. Empirical studies on PISA reading proficiency tests in Argentina (Debeer et al., 2017) and Japan (Okumura, 2014), and reading and mathematical competence tests in German fifth grade students students' (Pohl et al., 2014), found that performance was negatively related to their skipping propensity, and modeling reading proficiency with a IRTree vs. 1PL model resulted in systematic differences in parameter estimation. The tree model was the best fit to the

data, highlighting that missing responses in educational achievement data should not be ignored. Additionally, IRTrees have been successfully applied to online learning data by accounting for the speed-accuracy trade off in ability estimation (Coomans et al., 2016; Hofman et al., 2018), but this approach did not take problem-skipping into account.

1.3 The current study

To date, a hierarchical decision model, allowing problem-skipping and accuracy to be modeled as separate latent traits, has not been applied to large-scale adaptive learning data. Leveraging behavioral patterns in the form of skipped responses can give insight into learner meta-cognition and disengagement, and by providing a more accurate model of student ability can contribute to more fairness in adaptive learning systems.

In this work, we estimate the relationship between problem-skipping and accuracy in a large-scale adaptive educational software designed for primary school students to practice arithmetic. We compare unidimensional (1PL) and multidimensional IRTree models with varying levels of parameter constraints. These models allow us to examine whether responses to items are driven by one latent trait (ability) or whether accuracy and problem-skipping represent distinct underlying processes:

RQ1: Are problem-skipping and accuracy distinguishable processes in online learning?

Finding evidence for multidimensionality in ability assessment would have large implications for current educational measurement, which largely assume ability as a unidimensional construct. In addition, fitting the IRTree model to response data allows us to examine the latent correlations between nodes for items and students. This would help shed light on the currently sparse findings concerning the relationship between accuracy and problem-skipping:

RQ2: Given that problem-skipping and accuracy are distinct, how are they related?

2 Methods

2.1 Data

2.1.1 Online Learning Environment

We use data from an adaptive learning platform, [X] developed for primary school students to study Math and Language [1]. The platform currently consists of 65 games divided between environments designed for Arithmetic practice (28 games), language learning (24 games), and for second-language English learning (13 games). It is currently being used by over 450 000 children across 4300 schools. Depending on the difficulty of the task, the student can select one out of multiple answer options, or fill in the number themselves. They can skip the item by clicking a question mark. Then, the correct answer is shown and the student continues to the next item. While on the platform, children take care of a virtual garden, where each game represents a different learning domain, and the amount of practice is shown as the health of the plant.

In this work, we use data from a game called Series (Figure 2), wherein a student is asked to fill in a missing number in a number sequence. This game is designed to test knowledge of number sequences and relations. Each item that a student completes is given a score, determined by an explicit scoring rule (Figure 3, Equation 7). This scoring rule is visualized to the student in the form of coins: coins disappear for every second until the deadline (20 seconds). If the item is answered correctly, the amount of coins remaining is awarded, whereas if the answer is incorrect, the remaining amount of coins is lost. If the student skips the item or does not answer in time, no change is made to their score.

Item difficulties and student abilities are estimated with the Elo Rating System (ERS). When entering a learning game for the first time, a student is given a provisional ability rating θ (around the mean of the existing students within the same grade level). For each item that the student responds to, θ is updated according to the weighted difference between the accuracy and

[1] In the current version of this manuscript, the name of the learning platform, and the country in which it is active, is hidden for anonymity. In places where the final manuscript would refer to the platform, the name has been replaced with [X].

the expected accuracy on the item. Simultaneously, the rating of the item (β) is updated:

$$\hat{\theta}_p = \theta_p + K_p(S_{ip} - E(S_{ip})), \tag{5}$$

$$\hat{\beta}_i = \beta_i + K_i(E(S_{ip}) - S_{ip}). \tag{6}$$

K is a scaling parameter which determines the weight that the difference between the observed and expected score will have on the new ability estimate (for more information about the K-factor, see Vermeiren et al. (2024)). The score of person p following their response x on item i is dictated by the difference between their response time t and the deadline d, scaled by a discrimination parameter α:

$$S_{ip} = (2x_{ip} - 1)(\alpha_i d_i - \alpha_i t_i p). \tag{7}$$

This is the formalization of the explicit scoring rule (Figure 3). It is a so-called high speed, high stakes (HSHS) scoring rule (Maris & Van der Maas, 2012), which imposes a speed accuracy trade-off in response behavior. The discrimination parameter, α_i, is equal to $1/d$. In this way, the model maintains a constant discrimination value across items, and thus the calculated expected score is equivalent to a 1PL IRT model:

$$E(S_{ip}) = a_i d_i \frac{e^{2a_i d_i(\theta_p - \beta_i)} + 1}{e^{2a_i d_i(\theta_p - \beta_i)} - 1} - \frac{1}{\theta_p - \beta_i}. \tag{8}$$

The ERS adaptively matches children to items according to their desired difficulty level, which correspond to the probability of answering an item correctly. Difficulty levels are 60% (hard) 75% (medium), or 90% (easy). Children play on the default medium setting unless they select otherwise.

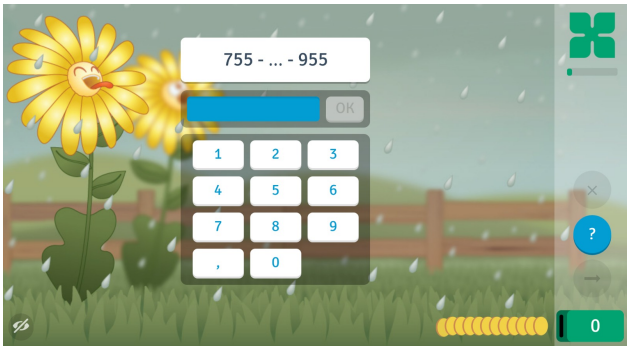

Figure 2: A screenshot of the Series game. A student is asked to fill in the missing number in the sequence. They gain (correct answer) or lose (incorrect answer) coins corresponding to the remaining amount of time in seconds (bottom right). A student can click the question mark (middle right) to retrieve the answer to the item and move on to the next.

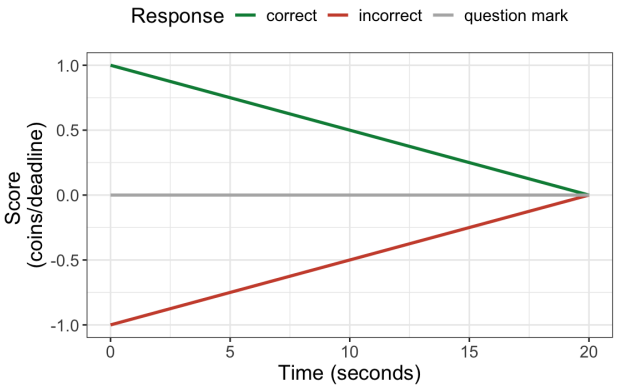

Figure 3: Visualization of the explicit scoring rule.

2.1.2 Inclusion criteria

To obtain data suitable for the IRT fitting procedure we selected a specific subset of data from the Series game. First, we sampled data from the 2-year period between 01–09–2014 and 31–08–2016. In recent years, changes have been made to the explicit scoring rule and the accessibility of the question mark response in [X]. To ensure that these changes do not affect question mark behavior in our analyses, we chose to look at data before this period of time. Lastly, the adaptive nature of the online learning environment means that user ability parameters are continuously updated. This also implies that for each user, there are periods of stable (learning is constant) and unstable (denoted by jumps or shifts in learning) measurements of user ability. In order to control for possible effects of learning on question mark usage, we only included data from periods where a user's ability remained relatively stable. We defined a stable period as one where the user's ability rating stayed close to their average rating. Specifically, a stable period was when the user's ability rating did not fluctuate more than 1 rating point above or below their average rating. The average rating for each user was calculated by taking the 20% trimmed mean of all their ability ratings, to prevent outliers from skewing the average. This criterion helps to minimize the influence of learning (or

sudden jumps in ability) on question mark usage. Lastly, for proper identification of the models, we selected only users who had skipped a minimum of 10 responses within their data period. The length of each obtained period of learning data varied between users, with the mode around 1–2 weeks and the max up to 80 weeks.

On the item side, we selected items which had been played a minimum of 2000 times within the selected data collection period. The final dataset consisted of 3795 respondents and 386 items.

2.2 Analyses

Comparing unidimensional and multidimensional linear response tree models allows to test whether one underlying trait gives rise to the response categories on a test item (De Boeck & Partchev, 2012). Here, we compare a unidimensional (1PL) model to IRTree models with varying parameter constraints to investigate whether responses to items in the Series game rely solely on one latent person variable (ability) or whether accuracy is distinguishable from problem-skipping. Because IRTree models are forms of generalized linear mixed models, the software package *lme4* (Bates et al., 2015) is used to estimate them. In all models, response is a binary variable denoting a decision (left or right) in the response tree. Each response in the tree is modeled as a 1PL model. The response tree is formulated such that node 1 represents the propensity to skip a problem, with 1 coded as a question mark response and 0 coded as an attempt to solve the item. Node 2 represents the propensity to give an incorrect response (difficulty), with 1 coded as an incorrect response and 0 coded as a correct response. See Figure 1 and Equations 3 and 4 for the technical specifications of the IRTree models. As such, each model is multilevel logistic regression which predicts response as a function of users and items, with varying constraints on the item- and user- specific random slopes:

1. Fully estimated IRTree: A multidimensional linear response tree where response is predicted by a random node effect of items and a random node effect of users. This model captures item- and user-specific variation across nodes.

2. Item-constrained IRTree: A multidimensional linear response tree where response is predicted by a random intercept for items, and a random node effect of users, but without modeling item-specific variation in nodes. This model captures variability in user behavior across nodes, but not item thresholds.

3. User-constrained IRTree: A multidimensional linear response tree where response is predicted by a random intercept for users, and a random node effect of items, but without modeling user-specific variation in nodes. This model captures variability in item thresholds across nodes, but not user behavior.

4. Fully constrained IRTree: A unidimensional random item model for linear response trees, where both item- and user-level effects are modeled with random intercepts, but without accounting for variation across nodes. Item omission and correctness are collapsed into the same response process, making this conceptually equivalent to a 1PL model.

3 Results

3.1 Description of Problem-Skipping

On average, children skipped approximately 6% of items in the Series game. Our exploration of problem-skipping behavior revealed several trends. First, unsurprisingly, the usage of the question mark feature was used more by children playing on the hard level ($M = 0.086, SD = 0.281$), compared to easy ($M = 0.043, SD = 0.203$) and medium ($M = 0.050, SD = 0.218$) levels. Additionally, there was a small increase in the overall use of question mark responses across grades (0.061 (SD = 0.239) for Grade 3, 0.059 (SD = 0.237) for Grade 4, 0.062 (SD = 0.241) for Grade 5, 0.062 (SD = 0.241) for Grade 6, 0.063 (SD = 0.244) for Grade 7, and 0.067 (SD = 0.250) for Grade 8.). However, when looking at average question mark usage across grades and difficulty levels simultaneously, this relationship changed, with stable proportions of problem-skipping for children playing on the hard level, and a decreasing proportion of problem-skipping across grades

for easy and medium levels (Figure 4).

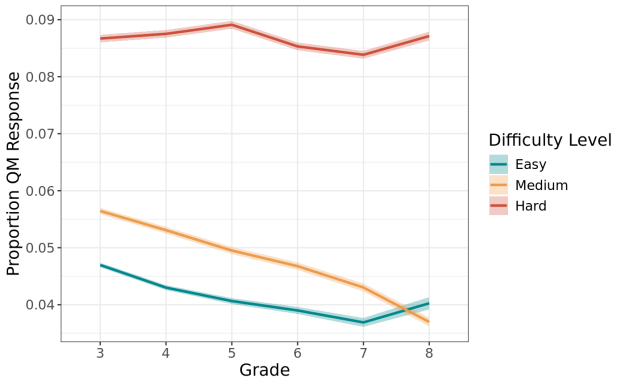

Figure 4: Proportion of question mark responses in the Series game, separated by grade and difficulty level (easy = 90% correct; medium = 75% correct; hard = 60% correct). Ribbons around the line denote the 95% confidence interval. Grade refers to the Dutch grade levels; age ranges approximately between 4 and 12 years. QM = question mark.

We also found differences with regards to at what point in time the question mark feature was used. Figure 5 displays the probability of each response type (correct, incorrect, question mark) across response times. Problem-skipping was much more likely to occur in the first five seconds upon the presentation of an item, and very unlikely to occur after five seconds, though with a small increase in probability upon nearing the response deadline at 20 seconds. This is in stark contrast to the probability of making a correct response, which quickly increased in the first five seconds and slowly decreased leading up to the deadline. The probability of making an incorrect response showed a small peak early in the response process, and slowly increased with increasing response times. This figure also demonstrates that within the response times that are most common (where the histogram bars are the highest), the probability of giving a question mark response is relatively low.

Similarly, with increasing number of items played in the Series game, the probability of skipping the problem decreased (Figure 6). This indicates that as users gain more understanding of how to play the Series game, their problem-skipping decreases. This graph also demonstrates the adaptivity of the system: as user ability estimates get more certain, the probability of making an error and question mark response stabilizes.

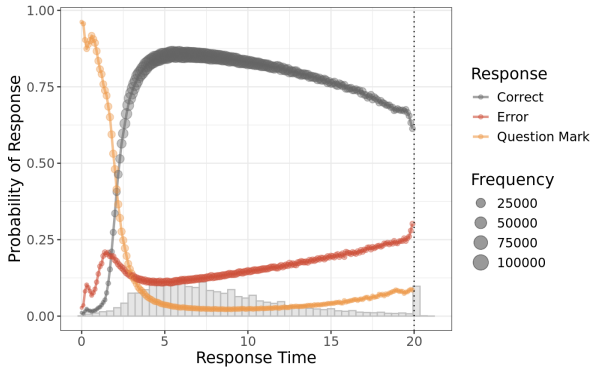

Figure 5: The probability of each response type (correct, error, question mark) across response times (seconds). Point size refers to the amount of observations at that data point. The vertical line denotes the deadline to respond, at 20 seconds. The histogram displays the average distribution of response times in the Series game.

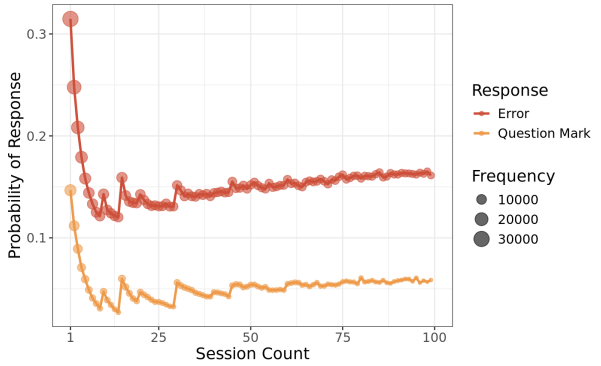

Figure 6: The probability of answering with a question mark or giving an incorrect response, over the first 100 items played in the Series game. Point size refers to the amount of observations at that data point.

3.2 Models of Problem-Skipping

To test whether problem-skipping is distinguishable from accuracy in the Series game, we compared a fully estimated multidimensional response tree to a unidimensional response tree, and two versions of a multidimensional response tree where either random item estimates (item-constrained IRTree) or random user estimates (user-constrained IRTree) were constrained (full model specifications are highlighted in Section 2.2). We compared these models on the basis of AIC and BIC model fit (Table 1). The fully estimated IRTree model fit the data best, supporting the inclusion of a latent trait for problem-skipping in the estimation of student ability.

The fully estimated IRTtree model had a fixed intercept of -1.85 ($p < .001$). This translates to a baseline probability of making an incorrect response of approximately 14.1%. Crucially, the correlation between the propensity to problem-skip (node 1) and give an incorrect response (node 2) was 0.41: students that were more likely to problem-skip were moderately more likely to make erroneous responses. Although this correlation is positive, it is not 1 (which is the assumption under the unidimensional model). Similarly, the correlation between the item skipping threshold and item difficulty was 0.77. That is, items that were more likely to be skipped were more likely

Model	Random Parameters	AIC	BIC	RMSE	cor($\theta^{(1)}, \theta^{(2)}$)	cor($\beta^{(1)}, \beta^{(2)}$)
1	$\theta_p, \theta_i, \beta_p, \beta_i$	1519007	1519095	0.3217	0.41	0.78
2	$\theta_p, \beta_p, \theta_i = \beta_i$	1533720	1533783	0.3234	0.35	-
3	$\theta_i, \beta_i, \theta_p = \beta_p$	1550148	1550211	0.3251	-	0.73
4	$\theta_i = \beta_i, \theta_p = \beta_p$	1566326	1566377	0.3268	-	-

Table 1: Model fit statistics and latent correlation estimates for the four fitted models. RMSE = root mean square error. RMSE values were computed via k-fold cross-validation, where 10% of the data was held out in each fold, the model was fit on the remaining data, and RMSE on the model predictions was calculated on the held-out set. The reported RMSE is the average across all folds.

to be estimated as difficult. The correlations between nodes for items and users is visualized in Figure 7.

3.3 Individual Patterns in Problem-Skipping

In addition to indicating the average relationship between individual propensities to problem-skip and give an incorrect response, the scatter plot of user estimates in Figure 7 reveals the spread of data around this relationship. From this, it could be of particular interest to look closer at the play behavior of students that either conform to or deviate from the average pattern of problem-skipping and accuracy. Exploring individual patterns in problem-skipping could reveal further insights and generate new hypotheses about individual differences in interacting with online learning systems. To explore this, we randomly selected users from 4 sets of criteria: (1) Low accuracy and high skipping rate (upper right corner of the scatter plot); (2) High accuracy and high skipping rate (lower right corner of the scatter plot); (3) High accuracy and low skipping rate (lower left corner of the scatter plot); (4) Low accuracy and low skipping rate (upper left corner of the scatter plot). We visualized their interactions with items in the Series game by plotting their response type (correct, incorrect, question mark) for each item across time, against their response time (Figure 8). This allows us to look at patterns in response types across time as well as whether responses were fast or slow and how this fluctuates.

This exploration revealed some interesting patterns. For example, because user ratings cannot improve with a question mark response (see Figure 3), users with a high ability should, in

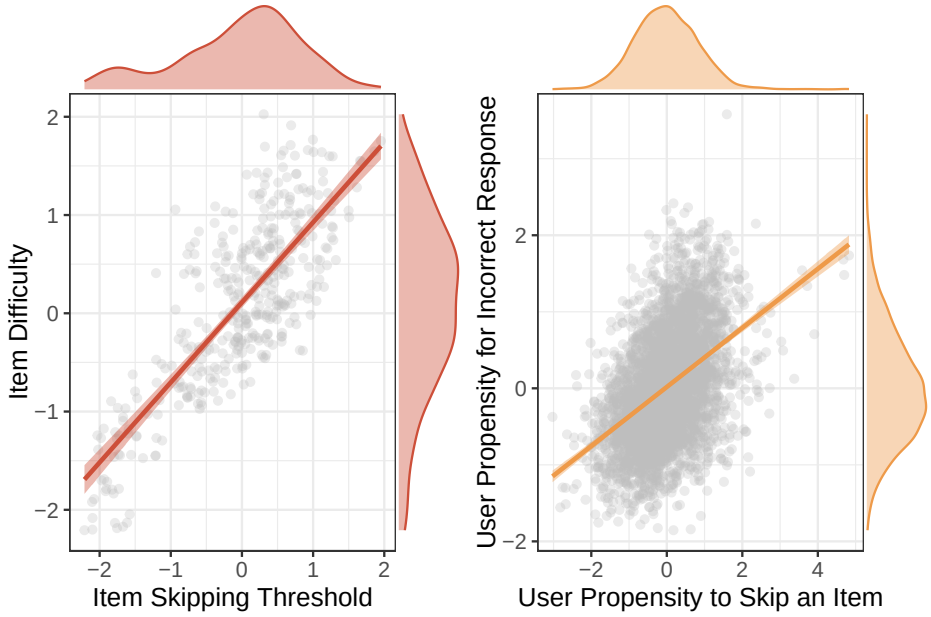

Figure 7: Correlation between estimated nodes in the fully estimated IRTree model. The left graph displays the correlation between the item skipping threshold ($\beta^{(1)}$), and the item difficulty ($\beta^{(2)}$). The right-hand graph displays the correlation between the individual propensity to skip an item ($\theta^{(1)}$), and the individual propensity to answer incorrectly ($\theta^{(2)}$). For both graphs, The scatter plot denotes individual data points, and the regression line displays the best-fitting linear relationship between the nodes. Density plots denote the distribution of each random parameter.

theory, not resort to using the question mark response at a high rate. When taking a closer look at these users (users C and D in Figure 8), they seem to be resorting to a different strategy compared to the other sampled users: long sequences of question-mark responding to avoid giving an answer, combined with sequences of attempting the item wherein the answer is mostly correct. In fact, the majority of users within this region of average skipping and accuracy showed such sequences, whereas the majority of users outside this region did not.

4 Discussion

Accurate educational assessment demands accurate measurement models. In this study, we compared a unidimensional IRT model, which assumes that problem-skipping does not influence student ability differently from accuracy, to a multidimensional IRTree model, which explicitly accounts for differences in propensity to skip an item. We found evidence that problem-skipping and ability

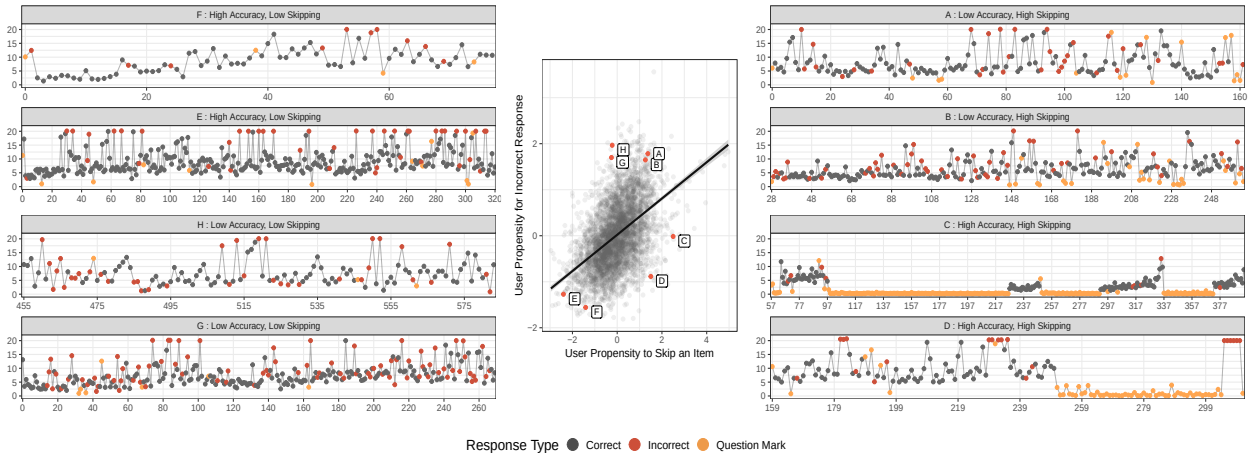

Figure 8: Eight individual users extracted based on their estimated propensity to skip and make an incorrect response (center graph). Individual interaction patterns are displayed in a clockwise fashion starting from the upper right corner (low accuracy, high skipping), to the upper left corner (low accuracy, low skipping) of the scatter plot. Each graph plots each each individual's response sequence to items in the Series game. The position on the y axis reflects the response time to the item. Colors denote the response type (correct, incorrect, question mark). While some users had longer response sequences in the extracted data (users C and E), the x axis is limited to showing the first 300 responses, to ease the visualization.

are distinguishable processes, and as such, should not be measured on the same scale. Additionally, we found a moderate, positive association between problem-skipping and accuracy, suggesting that students with a high propensity to skip problems are somewhat more likely to be inaccurate when they do give a response. However, the relative weakness of this correlation opposes the assumption of unidimensional measurement models that these are perfectly correlated. Instead, it implies that skipped responses carry information about both accuracy and other meta-cognitive or motivational traits which affect responding. Failure to account for this will introduce bias to measurement and limit the adaptivity of the system.

Our findings extend previous work on response scale- and achievement test data—demonstrating that ignoring skipped items may lead to bias in the measurement of latent traits (Debeer et al., 2017; Holman & Glas, 2005; Huang, 2020; Okumura, 2014; Pohl et al., 2014)—to online learning platforms. They also shed some light on the mixed evidence concerning the relationship between problem-skipping and accuracy, adding to the findings of Norum et al. (2024) that more problem-skipping is associated with more errors, and the body of literature indicating individual differences

in the use of alternative response options (Goldin et al., 2012, 2013; Norum et al., 2024). However, the evidence that problem-skipping and accuracy are distinct and only moderately related may also help to explain the discord in the literature. Specifically, the imperfect relationship suggests that the observed association between problem-skipping and accuracy may be influenced by contextual factors, such as the design of the learning environment or task difficulty, which give space for different manifestations of meta-cognitive strategies and motivational states such as frustration or boredom. Future work should aim to disentangle these contextual influences to better understand the role of problem-skipping in learning and assessment.

Many adaptive learning platforms include a response option alternative to providing a direct answer to an item, such as skipping or requesting a hint. These findings therefore have widespread relevance for any adaptive learning system that aims to provide a fair estimate of student ability while allowing flexibility in how students engage with items. A basic recommendation for such systems is to, at a minimum, measure both latent traits once the data have been collected. The IRTree model has proven a feasible method for this kind of post-hoc analysis, offering insights into individual differences in both accuracy and problem-skipping. However, there are some limitations to the use of IRTree models: they require large amounts of data for reliable estimates, making them less suitable for students who interact with the learning platform less frequently. Moreover, in the way that it is currently implemented, this method reflects past behavior rather than providing real-time insights, a core feature in adaptive learning systems. In practice, the IRTree framework could be used to incorporate a second parameter into the measurement model for real-time tracking of problem-skipping. This would enable adaptive systems to monitor and act on these tendencies as they emerge, for example via teacher dashboards. While this method is still data-intensive, it would allow for dynamic tracking of both accuracy and problem-skipping, providing a fuller picture of the learning process.

Despite their importance in revealing individual differences in how problem-skipping relates to accuracy, post-hoc and real-time models reveal little about where these differences stem from. This raises the question of whether different types of meta-cognitive strategies or states–such as

boredom, frustration, gaming, or wheel-spinning–can be inferred from students' response data. Consider two students, one of whom quickly skips items in a very long sequence (students C and D, Figure 8) to avoid effort, and another who is exerting effort into learning by attempting as many items as they can, but skips the item as soon as they are uncertain to avoid the risk of negative feedback. While both students would display high problem-skipping tendencies, they bode for different intervention styles and may therefore want to be separated. To aid in this, more information can be added to the tree model, such as a node to distinguish fast from slow responses (Coomans et al., 2016; Hofman et al., 2018). An IRTree model can also be extended to model different levels of disengagement, by, for example, explicitly accounting for drop-out (Huang, 2020).

Apart from merely measuring different cognitive and meta-cognitive states from observed response data, adaptive systems can be designed to make these strategies directly observable. One approach is to provide response options explicitly linked to meta-cognitive strategies, allowing systems to infer and adapt to the strategies students employ. The feasibility of such an approach would depend on the scale and flexibility of the learning system. For instance, in a system with an exploratory style of learning, mapping response options to every possible strategy may prove challenging. An alternative approach is to change the scoring rule in such a way that it constrains students' behaviors to a specific strategy. For example, in [X], the HSHS scoring rule (Maris & Van der Maas, 2012) is implemented in such a way that students adhere to a speed-accuracy trade-off. In a similar manner, a separate scoring rule for problem-skipping can be implemented such that it has an appropriate trade-off with accuracy. A potential future research avenue is to investigate the applicability and validity of such approaches for detecting different problem-skipping strategies.

Lastly, if problem-skipping is seen as undesirable behavior–such as indicating low effort or gaming the system–the design of the system or scoring rule can be changed in such a way that this behavior is decreased. Such an intervention has been successfully implemented, wherein the ability to use the problem-skipping option was delayed upon the presentation of each item (Savi et al., 2018). This led to an increase in the amount of items that students attempted, promoting

effortful responding in the system. Such interventions highlight the potential of system design to nudge students toward more productive behaviors.

A final note to consider is that this model assumes that the student undergoes a sequential decision process, *i.e.*, first deciding whether to respond to a problem or skip it, and only then deciding on the actual response (if not skipped). It is not certain that this is the true decision process of the student. In the future, it would be interesting to compare different modeling frameworks aimed at deciphering the response process of students. For example, Race Models (Heathcote & Matzke, 2022) assume that there is competition between two or multiple parallel processes, and as enough evidence accumulates in support of one process, this process wins and determines the subsequent decision. In the context of problem-skipping, such a model may be suitable for detecting fine-grained processes such as individual differences in processing speed, especially in the context of a speed accuracy trade-off (Evans et al., 2018; Heathcote & Matzke, 2022). Alternative methods to model students' decision-making processes include Bayesian Hierarchical Models (Lee & Wagenmakers, 2014; Sawyer et al., 2018) and Markov Decision Processes (LaMar, 2018; Puterman, 1994). The rich, granular data generated by online learning systems offer an ideal opportunity to compare cognitive decision models, which usually require detailed and structured data collected in lab experiments. Nevertheless, the current results already indicate that there are structural individual differences which, regardless of how the decision procedure is modeled, needs to be accounted for.

5 Conclusion

In a large-scale adaptive learning environment for arithmetic practice, we found that accuracy and problem-skipping are moderately related, but distinct traits. Under a measurement model which doesn't explicitly account for individual differences in problem-skipping, student ability estimates could be biased towards children who respond differently to problem-skipping options. To prevent this, online learning environments should consider estimating and reporting latent traits for both

the tendency to problem-skip and give an accurate response, or intervene in the learning system such that strategies can be inferred or problem-skipping is reduced. We have shown the feasibility of applying the IRTree framework (De Boeck & Partchev, 2012) to model problem-skipping in an adaptive learning platform. This approach has the potential to be extended to model other meta-cognitive strategies, or model problem-skipping in domains apart from mathematics. Only with accurate measurement models can equity in learning analytics be realized.

6 Data and Code Availability

The dataset used in the current study was collected under license of the third party company to which the data belongs. Restrictions apply to the availability of these data, but can be made available from the authors upon reasonable request and with permission of the company. Source code used to fit and analyze the data is publicly available at https://anonymous.4open.science/r/qm-trees-10C5/.

References

Aleven, V., & Koedinger, K. R. (2000). Limitations of student control: Do students know when they need help? *Intelligent Tutoring Systems*, 292–303.

Baker, R. S., Corbett, A. T., & Aleven, V. (2008). More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian Knowledge Tracing [Series Title: Lecture Notes in Computer Science]. In *Intelligent Tutoring Systems* (pp. 406–415, Vol. 5091). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-69132-7_44

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004). Detecting student misuse of intelligent tutoring systems. *Intelligent Tutoring Systems: 7th International Conference, ITS 2004, Maceió, Alagoas, Brazil, August 30-September 3, 2004. Proceedings 7*, 531–540.

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the cognitive tutor classroom: When students "game the system". *Proceedings of the SIGCHI*

conference on Human factors in computing systems, 383–390. https://doi.org/https://doi.org/10.1145/985692.985741

Baker, R. S., Roll, I., Corbett, A. T., & Koedinger, K. R. (2005). Do performance goals lead students to game the system? *AIED*, 57–64.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beck, J. E. (2004). Using response times to model student disengagement. *Proceedings of the ITS2004 Workshop on Social and Emotional Intelligence in Learning Environments*, *20*(2004), 88–95.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological methods*, *17*(4), 665. https://doi.org/https://doi.org/10.1037/a0028111

Colantonio, J. A., Bascandziev, I., Theobald, M., Brod, G., & Bonawitz, E. (2024). Predicting Learning: Understanding the Role of Executive Functions in Children's Belief Revision Using Bayesian Models. *Topics in Cognitive Science*, tops.12749. https://doi.org/10.1111/tops.12749

Conati, C., Jaques, N., & Muir, M. (2013). Understanding Attention to Adaptive Hints in Educational Games: An Eye-Tracking Study. *International Journal of Artificial Intelligence in Education*, *23*(1-4), 136–161. https://doi.org/10.1007/s40593-013-0002-8

Coomans, F., Hofman, A., Brinkhuis, M., van der Maas, H. L., & Maris, G. (2016). Distinguishing fast and slow processes in accuracy-response time data. *PloS one*, *11*(5), e0155149. https://doi.org/https://doi.org/10.1371/journal.pone.0155149

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modelling and User-Adapted Interaction*, *4*(4), 253–278. https://doi.org/10.1007/BF01099821

Corrigan, S., Barkley, T., & Pardos, Z. (2015). Dynamic approaches to modeling student affect and its changing role in learning and performance. *User Modeling, Adaptation and Personalization*, 92–103.

De Boeck, P., & Partchev, I. (2012). Irtrees: Tree-based item response models of the glmm family. *Journal of Statistical Software*, *48*, 1–28. https://doi.org/https://doi.org/10.18637/jss.v048.c01

Debeer, D., Janssen, R., & De Boeck, P. (2017). Modeling skipped and not-reached items using irtrees. *Journal of Educational Measurement*, *54*(3), 333–363. https://doi.org/https://doi.org/10.1111/jedm.12147

Deonovic, B., Yudelson, M., Bolsinova, M., Attali, M., & Maris, G. (2018). Learning meets assessment: On the relation between item response theory and Bayesian knowledge tracing. *Behaviormetrika*, *45*(2), 457–474. https://doi.org/10.1007/s41237-018-0070-z

Evans, N. J., Steyvers, M., & Brown, S. D. (2018). Modeling the covariance structure of complex datasets using cognitive models: An application to individual differences and the heritability of cognitive ability. *Cognitive science*, *42*(6), 1925–1944. https://doi.org/https://doi.org/10.1111/cogs.12627

Fidan, A., & Koçak Usluel, Y. (2024). Emotions, metacognition and online learning readiness are powerful predictors of online student engagement: A moderated mediation analysis. *Education and Information Technologies*, *29*(1), 459–481. https://doi.org/10.1007/s10639-023-12259-6

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*(3), 225–245. https://doi.org/https://doi.org/10.1111/j.1745-3984.2008.00062.x

Fleur, D. S., Bredeweg, B., & Van Den Bos, W. (2021). Metacognition: Ideas and insights from neuro- and educational sciences. *npj Science of Learning*, *6*(1), 13. https://doi.org/10.1038/s41539-021-00089-5

Goldin, I. M., Koedinger, K. R., & Aleven, V. (2012). Learner differences in hint processing. *International Educational Data Mining Society*.

Goldin, I. M., Koedinger, K. R., & Aleven, V. (2013). Hints: You can't have just one. *Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013)*.

Gorgun, G., & Bulut, O. (2022). Considering disengaged responses in bayesian and deep knowledge tracing. *International Conference on Artificial Intelligence in Education*, 591–594.

Heathcote, A., & Matzke, D. (2022). Winner takes all! what are race models, and why and how should psychologists use them? *Current Directions in Psychological Science*, *31*(5), 383–394. https://doi.org/https://doi.org/10.1177/09637214221095852

Hofman, A. D., Visser, I., Jansen, B. R., Marsman, M., & van der Maas, H. L. (2018). Fast and slow strategies in multiplication. *Learning and Individual Differences*, *68*, 30–40.

Holman, R., & Glas, C. A. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*(1), 1–17. https://doi.org/https://doi.org/10.1111/j.2044-8317.2005.tb00312.x

Huang, H.-Y. (2020). A mixture irtree model for performance decline and nonignorable missing data. *Educational and Psychological Measurement*, *80*(6), 1168–1195. https://doi.org/https://doi.org/10.1177/0013164420914711

Karlen, Y. (2016). Differences in students' metacognitive strategy knowledge, motivation, and strategy use: A typology of self-regulated learners. *The Journal of Educational Research*, *109*(3), 253–265. https://doi.org/10.1080/00220671.2014.942895

Kim, B., Park, H., & Baek, Y. (2009). Not just fun, but serious strategies: Using meta-cognitive strategies in game-based learning. *Computers & Education*, *52*(4), 800–810. https://doi.org/10.1016/j.compedu.2008.12.004

LaMar, M. M. (2018). Markov decision process measurement model. *Psychometrika*, *83*(1), 67–88. https://doi.org/https://doi.org/10.1007/s11336-017-9570-0

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge university press.

Maris, G., & Van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*(4), 615–633. https://doi.org/https://doi.org/10.1007/s11336-012-9288-y

Matthews, G., Hillyard, E. J., & Campbell, S. E. (1999). Metacognition and maladaptive coping as components of test anxiety. *Clinical Psychology & Psychotherapy*, *6*(2), 111–125. https://doi.org/https://doi.org/10.1002/(SICI)1099-0879(199905)6:2<111::AID-CPP192>3.0.CO;2-4

Mislevy, R. J., & Wu, P.-K. (1996). Missing responses and irt ability estimation: Omits, choice, time limits, and adaptive testing. *ETS Research Report Series*, *1996*(2), i–36. https://doi.org/https://doi.org/10.1002/j.2333-8504.1996.tb01708.x

Norum, R., Lee, J.-E., Ottmar, E., & Harrison, L. (2024). Student profiles based on in-game performance and help-seeking behaviours in an online mathematics game. *British Journal of Educational Technology*, bjet.13463. https://doi.org/10.1111/bjet.13463

Okumura, T. (2014). Empirical differences in omission tendency and reading ability in pisa: An application of tree-based item response models. *Educational and Psychological Measurement*, *74*(4), 611–626. https://doi.org/https://doi.org/10.1177/0013164413516976

Ostrow, K., Donnelly, C., Adjei, S., & Heffernan, N. (2015). Improving student modeling through partial credit and problem difficulty. *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*, 11–20. https://doi.org/10.1145/2724660.2724667

Pekrun, R., & Perry, R. P. (2013, July). Control-value theory of achievement emotions. In *International Handbook of Emotions in Education*. Routledge. https://doi.org/10.4324/9780203148211.ch7

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement*, *74*(3), 423–452. https://doi.org/https://doi.org/10.1177/0013164413504926

Puterman, M. L. (1994). *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons.

Rabinowitz, M. (2017). The Interaction Between Knowledge, Strategies, Metacognition, and Motivation. In *Psychology of Learning and Motivation* (pp. 35–52, Vol. 67). Elsevier. https://doi.org/10.1016/bs.plm.2017.03.002

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests.* ERIC.

Šarić-Grgić, I., Grubišić, A., & Gašpar, A. (2024). Twenty-five years of Bayesian knowledge tracing: A systematic review. *User Modeling and User-Adapted Interaction*, *34*(4), 1127–1173. https://doi.org/10.1007/s11257-023-09389-4

Savi, A. O., Ruijs, N. M., Maris, G. K., & van der Maas, H. L. (2018). Delaying access to a problem-skipping option increases effortful practice: Application of an a/b test in large-scale online learning. *Computers & Education*, *119*, 84–94. https://doi.org/https://doi.org/10.1016/j.compedu.2017.12.008

Savi, A. O., van Klaveren, C., & Cornelisz, I. (2023). Combating effort avoidance in computer adaptive practicing: Does a problem-skipping restriction promote learning? *Computers & Education*, *206*, 104908. https://doi.org/https://doi.org/10.1016/j.compedu.2023.104908

Sawyer, R., Rowe, J., Azevedo, R., & Lester, J. (2018). Modeling player engagement with bayesian hierarchical models. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, *14*(1), 257–263. https://doi.org/https://doi.org/10.1609/aiide.v14i1.13048

Shields, M., Calabro, G., & Selmeczy, D. (2024). Active help-seeking and metacognition interact in supporting children's retention of science facts. *Journal of Experimental Child Psychology*, *237*, 105772. https://doi.org/10.1016/j.jecp.2023.105772

Sungur, S. (2007). Modeling the relationships among atudents' motivational beliefs, metacognitive strategy use, and effort regulation. *Scandinavian Journal of Educational Research*, *51*(3), 315–326. https://doi.org/10.1080/00313830701356166

Taub, M., Azevedo, R., Bouchet, F., & Khosravifar, B. (2014). Can the use of cognitive and metacognitive self-regulated learning strategies be predicted by learners' levels of prior

knowledge in hypermedia-learning environments? *Computers in Human Behavior*, *39*, 356–367. https://doi.org/10.1016/j.chb.2014.07.018

Thissen-Roe, A., & Thissen, D. (2013). A two-decision model for responses to likert-type items. *Journal of Educational and Behavioral Statistics*, *38*(5), 522–547. https://doi.org/https://doi.org/10.3102/1076998613481500

Vermeiren, H., Hofman, A. D., Bolsinova, M., van der Maas, H. L., & Van Den Noortgate, W. (2024). Balancing stability and flexibility: Investigating a dynamic k value approach for the elo rating system in adaptive learning environments [Preprint]. *OSF*. https://doi.org/https://doi.org/10.31234/osf.io/6hzke

Wang, Y., & Heffernan, N. (2013). Extending knowledge tracing to allow partial credit: Using continuous versus binary nodes. In *Artificial Intelligence in Education* (pp. 181–188, Vol. 7926). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-39112-5_19

Wang, Y., Heffernan, N. T., & Beck, J. E. (2010). Representing student performance with partial credit.

Wells, A. (1995). Meta-cognition and worry: A cognitive model of generalized anxiety disorder. *Behavioural and Cognitive Psychotherapy*, *23*(3), 301–320. https://doi.org/10.1017/S1352465800015897

Winne, P. H. (2017, September). Cognition and metacognition within self-regulated learning. In *Handbook of Self-Regulation of Learning and Performance* (2nd ed., pp. 36–48). Routledge. https://doi.org/10.4324/9781315697048-3

Zhao, L., & Ye, C. (2020). Time and performance in online learning: Applying the theoretical perspective of metacognition. *Decision Sciences Journal of Innovative Education*, *18*(3), 435–455. https://doi.org/10.1111/dsji.12216