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

Annie Johansson* Alexander O. Savi† Abe D. Hofman‡

Preprint, November 2024

Abstract

The estimation of student ability is paramount in large-scale personalized learning. To this end, state-of-the-art adaptive learning environments use item response theory (IRT). Previous work in traditional learning assessment has demonstrated that unidimensional IRT models fall short in adequately estimating ability when items on a test are skipped. In this study, we extend this work to online learning platforms. We analyze data from a large-scale online learning platform used to practice Arithmetic and Language. Using the IRTree framework, we compare the unidimensional model of accuracy to a multidimensional model which additionally accounts for the decision to respond or skip a problem. We found support for problem-skipping as a non-ignorable process: students that were more likely to problem-skip were more likely to make erroneous responses. Further exploration revealed individual differences in the strategies involved with problem-skipping. To ensure that learning analytic tools are supported by fair measurement models, we suggest several ways to account for problem-skipping when estimating student ability.

Keywords

Problem-skipping, Item Response Theory, Item Response Trees, Adaptive Learning, Learning Analytics, Ability Estimation

*University of Amsterdam, Amsterdam, The Netherlands, a.m.johansson2@uva.nl
†University of Amsterdam, Amsterdam, The Netherlands
‡University of Amsterdam, Amsterdam, The Netherlands, Prowise Learn, Amsterdam, The Netherlands

1 Introduction

Accurately assessing and improving children's educational performance relies on our ability to effectively measure their cognitive processes during learning and testing. Over the years, the development of the field of learning analytics (LA) has been imperative for understanding student responding in education [40]. LA typically involves analyzing large datasets from digital learning environments, where student abilities are inferred through measurement models. These models often focus on accuracy as the primary latent trait driving performance. However, many online learning platforms offer response options beyond simply giving the answer, such as requesting hints or skipping questions (hereby referred to as problem-skipping). The latent traits underlying problem-skipping remain unclear, and they are often either ignored or treated as incorrect in traditional models. This study explores whether skipped responses represent a distinct cognitive process, separate from accuracy, in estimating learner ability.

2 Related Work

Item response theory (IRT) models are commonly used measurement models for assessing student ability in the context of educational practice and testing. In this section, we give a brief overview of how the standard IRT models and an extension, in the form of IRT Tree models, are formulated, and how alternative responses have been accounted for in previous work.

2.1 Classical IRT models

In the standard IRT 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. Common properties include item difficulty, which represents at what ability level the student has a 50% chance of responding to the item correctly, and item discrimination, indicating how well the item discriminates between students of different abilities. The 1-Parameter Logistic Model (1PL; also referred to as a Rasch Model [33]) considers item difficulty, and a 2-Parameter Logistic Model (2PL) considers both item difficulty and item discrimination. In this paper, we use extensions of the 1PL model to account for hierarchical decision processes, and thus do not proceed further with

explaining the 2PL model. In the 1PL model, the probability that a student with ability θ answers item i correctly is given by the function:

$$P(X_i = 1 \mid \theta) = \frac{1}{1 + e^{-(\theta - b_i)}} \tag{1}$$

where b_i is the difficulty parameter for item i.

In classic IRT models, skipped responses are often ignored or treated as incorrect. The choice to treat a response as missing and ignore it relies on the assumption that it is missing at random (MAR), i.e., the process of skipping a response is entirely random. When the assumption of MAR holds, the probability of the missing data pattern does not depend on the observed data, and the latent trait of the missing data process is unrelated to the estimated parameter in the data, such as the ability [34]. In such cases, the missing data is ignorable. Similarly, the choice to treat an alternative response as incorrect relies on the assumption that the probability of making such a response can be fully attributed to the estimated parameter in the data (for example, the process of skipping a problem is the same as the process of making a problem incorrectly). However, if in any case one of these assumptions are violated, that is, if the data is missing not at random (MNAR), omitted or missed responses are not ignorable and it is necessary to account for the missing data process in addition to the main parameters [26]

2.2 IRTree models

Previous work has demonstrated that ignoring missing responses in IRT approaches can lead to bias on person- and item-parameter estimates [13, 28, 31, 18, 11]. In classical test theory, it has been widely found that test-takers may exhibit undesirable response patterns which do not reflect the trait that the scale aims to measure. For example, on a five-point Likert scale, some respondents may consistently agree or disagree with the item by consistently choosing 1 or 5 (and ignoring 2 and 4), referred to as extreme response styles, and some may exhibit a stronger tendency to choose the middle option (3; mid-scale responding). 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 trait to be measured. Therefore, a range of research is dedicated to reliably identifying

and controlling for response styles to produce test results which are as unbiased as possible. One such way is to model the response process as a hierarchical decision tree, where the content and the response format of the item are both assumed to contribute to the observed response [3, 8, 38]. Following the previous example, the respondent may choose whether to endorse the item in any direction (mid-scale responding or not), and if they do, whether they agree or disagree. IRTree models have been used successfully for capturing extreme response styles and mid-scale responding on Likert-type response scales [4, 38].

IRTree models can also be used to model missing responses in an educational context. Here, responses are often categorized as correct, incorrect, or missing (i.e., a student skipped the item), and the problem-skipping process is modeled separately from accuracy (incorrect or correct). 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

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

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

Each node in the tree model is modeled with a 1PL model [33]. Thus, the tree model combined 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:

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

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} comprise a mapping matrix and together decide whether a node will contribute to the probability of category m. 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.

The main advantage with this modeling approach is that the unidimensionality of the decision process can be formally tested [8]. 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 (item skipping threshold), and
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 omissions depend on respondents or on items? The unidimensional assumes that these correlations are negligible, i.e., for individuals, problem-skipping and ability are the same, and the likelihood for an item to be skipped is equal to its estimated difficulty. In Rubin's [34] terms, the IRTree model becomes a model for MNAR which can be tested against the 1PL model which assumes MAR. If correlations between omission tendency and ability are not 1, and the tree model fits the data better, there is evidence that the missing data is not ignorable and needs to be accounted for in the estimation of student ability.

Previous work has found support for problem-skipping and ability as distinct processes in educational assessment data [29, 31, 11, 19, 18]. In a simulation study, Debeer et al. [11] found that a 1PL model results in bias in person- and item-parameters when data is MNAR, and that modeling the missingness mechanism with an IRTree model significantly reduces this bias. This was especially the case for students who omit many responses. In a subsequent empirical study on the PISA 2009 reading proficiency tests in Argentina, the authors found that students' reading proficiency was negatively related to their skipping propensity, and that 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, and, consequently, they conclude that missing responses in educational achievement data should not be ignored. Okumura [29] found similar results for Japanese students on the PISA 2009 reading proficiency tests. Similarly, Pohl et al. [31] compared a unidimensional model with a multidimensional model accounting for missing propensity on reading and

mathematical competence tests for German fifth grade students. They found that for both reading and mathematics, lower-ability students omitted more responses than higher-ability students. The unidimensional model was not suitable for estimating student ability, especially when omitted responses were treated as incorrect. When missing responses were ignored, a unidimensional model was sufficient to estimate student ability. Nevertheless, given the relationship between problem-skipping and ability, the authors argued that the missing mechanism is nonignorable, and should be accounted for in estimating student ability within the IRT framework.

2.3 Item omission in online learning

Aforementioned research on the relationship between omission tendencies and student ability has focused on traditional pen-and-paper testing. However, online learning environments (OLEs) introduce new dimensions to this relationship by offering response options that are not typically available in standardized tests. These options include skipping a question entirely, requesting hints, receiving partial answers, or accessing the full solution. Such features offer students greater flexibility and may encourage different response styles. Despite this added complexity, many OLEs rely on traditional IRT models to estimate student ability, and often ignore omitted responses or treat them as incorrect.

A key distinction between pen-and-paper assessments and online learning practice is the increased autonomy that students experience in OLEs. While standardized tests like PISA are designed to be low-stakes, students may still perceive them as formal and make a concerted effort to answer each question to the best of their ability. Additionally, these tests are typically short-term, meaning that most students can sustain a consistent level of attention, motivation, and effort throughout the test. In contrast, OLEs allow students to engage with content over multiple sessions, and their motivation or willingness to engage deeply may vary across different log-in occasions. Further, some students may learn to "game the system" [1] in online learning platforms, finding response strategies that maximize reward at the cost of learning. Although students are undoubtedly likely to show fluctuations in their motivation and effort throughout an achievement test, the free-practice nature of OLEs invokes many different situational and individual factors that influence a student's propensity to problem-skip.

On the other hand, when a student chooses to problem-skip in an OLE it is a deliberate act [35, 36], as opposed to in traditional settings where it can only be assumed that the student intentionally omitted the answer. If one would assume that a student chooses to skip an item because they know that they do not know the answer, it would be acceptable to treat the item omission as an incorrect response, and the measurement model likely would not produce bias.

2.4 The current study

Given the unique conditions of OLEs on the learning process, it is important to measure problem-skipping and its relation to student proficiency in this setting. The large-scale nature and richness of data provided by OLEs offer a unique opportunity to investigate this in an approachable manner. Previously, IRTrees have been successfully applied to online learning data by accounting for the speed-accuracy trade off in ability estimation [7].

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 fit different unidimensional and multidimensional IRT models and assess their ability to estimate student ability in the presence or absence of a latent trait for problem-skipping. Our research questions are as follows:

1. Are problem-skipping and accuracy distinguishable processes in online learning?
2. If so, what is the relationship between these two constructs?

3 Methods

3.1 Data

3.1.1 Online Learning Environment

We use data from the adaptive OLE Prowise Learn (previously known in the literature, and referred to here as, Math Garden [23]). The learning platform was developed for Dutch primary school students to study Math and Language. The platform currently consists of 65 games divided between the learning environments Math Garden (for Arithmetic practice; 28 games), Language Sea (for

Dutch language learning; 24 games), and Words and Birds (for English language 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. More information about the platform is available at https://www.prowise.com/en/product/prowise-learn/the Prowise Learn Website. Research on the adaptivity of the platform can be found in [23, 6, 5, 17]. The platform has also been used for applied developmental and educational research, across topics such as skill development in mathematics [30] and language [22], math anxiety [20, 21, 16], error-monitoring [10, 9], effort [36] and problem-skipping [35].

In this work, we use data from the Math Garden game 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 5). 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 the expected accuracy on the item. Simultaneously, the rating of the item (β) is updated:

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

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

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 [39]). 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:

$$S_{ip} = (2x_{ip} - 1)(\alpha_i d_i - \alpha_i t_i p). \tag{5}$$

This is the formalization of the explicit scoring rule (Figure 3). It is a so-called high speed, high stakes (HSHS) scoring rule [27], which imposes a speed accuracy trade-off in response behavior. In Math Garden, 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:

$$E(S_{ip}) = a_i d_i \frac{e^{2a_i d_i(\theta_p - \beta_i)} + 1}{e^{2a_i d_i(\theta_p - \beta_i)} - 1} - \frac{1}{\theta_p - \beta_i}. \tag{6}$$

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. For an in-depth explanation of how the ERS has been applied to Math Garden, see [23].

3.1.2 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 Math Garden. 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. Next, for identification of the IRTree model, we included only users who have made at least 20 question mark responses. Lastly, the adaptive nature of the OLE 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. 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 4110 respondents and 386 items.

3.2 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 [8]. 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 ability is distinguishable from problem-skipping. Because IRTree models are forms of generalized linear mixed models (GLMM), the software package *lme4* [2] 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. We compare the following models:

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.

4 Results

4.1 Description of Problem-Skipping Behavior

On average, children skipped approximately 6% of items in the Series game. Upon exploring problem-skipping behavior, we found 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).

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.

4.2 Models of Problem-Skipping Behavior

To test whether problem-skipping is distinguishable from ability 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 3.2). We compared these models on the basis of AIC and BIC model fit statistics (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 13.6%. Crucially, the correlation between the propensity to problem-skip (node 1) and give an incorrect response (node 2) was 0.44: students that were more likely to problem-skip were 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 to be estimated as difficult. The correlations between nodes for items and users is visualized in Figure 7.

Model	Random Parameters	AIC	BIC	cor($\theta^{(1)}, \theta^{(2)}$)	cor($\beta^{(1)}, \beta^{(2)}$)
Fully estimated IRTree	$\theta_p, \theta_i, \beta_p, \beta_i$	1559132	1559220	0.44	0.77
Item-constrained IRTree	θ_p, β_p	1574386	1574449	0.37	-
User-constrained IRTree	θ_i, β_i	1591531	1591594	-	0.72
Fully constrained IRTree	-	1619044	1619081	-	-

Table 1: Model fit statistics and latent correlation estimates for the four fitted models.

4.3 Differential Patterns of Problem-Skipping Behavior

Apart from 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 item ratings cannot improve with a question mark response (see Figure 3), users with a high ability should, in 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 ability showed such sequences, whereas the majority of users outside this region did not.

5 Discussion

Accurate educational assessment demands accurate measurement models. 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 are distinguishable processes, and that those who skip more items tend to be estimated with a lower ability. As such, IRT models aiming to estimate student ability need to account for problem-skipping. These 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 [18, 31, 19, 29, 11]—to online learning platforms. Many online learning platforms include an option to delay or omit an actual response, such as the option to skip the problem, request a hint, or another way of seeking help. These findings are therefore relevant for educators, developers, and researchers looking to estimate student ability with these tools.

There are several different ways to improve educational assessment by accounting for problem-skipping. The first, and simplest, option, is to measure both latent traits after the data have come in, and account for potential differences in ability and propensity to skip an item. The second option is to incorporate a latent trait measurement for problem-skipping into the learning algorithm. Similarly to Klinkenberg et al. [23]'s solution of using Elo-estimates for "on the fly" measurement of student ability, a second parameter relating to problem-skipping can be incorporated and updated as data come in. In this way, tools such as teacher dashboards which visualize the learning trajectories of students can reveal insights both into students' ability ratings in the current moment, as well as their skipping tendencies, and one can choose to intervene on either one of them. Lastly, if problem-skipping is assumed to be undesirable behavior (indicating low effort, gaming the system, etc.), one option is to alter the design of the learning platform or scoring model such that this behavior is decreased. Savi et al. [35] successfully implemented an intervention aimed at reducing effortless problem-skipping in Math Garden, by briefly delaying students' access to a problem-skipping option. Following the intervention, children increased the amount of items they attempted, which is ultimately what is required for active learning of new problems.

One limitation to the current approach is that it may not encompass all types of behaviors

that are important for distinguishing uncertainty from accuracy. Consider the case of one student, similar to students C and D in Figure 8, who quickly skips items in very long sequences. It may be that such a student is in an unmotivated state, trying to exert as little effort as possible. Thus, the student answers only items of which they are certain, and quickly skips items of which they are uncertain. In contrast, a student in a motivated state who aims to exert effort into the learning process may be attempting as many items as they can, and for uncertain items take time to think about the answer before resorting to problem skipping. The behavior of the unmotivated student is one which educators may want to pinpoint and intervene on, while the behavior of the motivated student is considered productive and should be encouraged. To model this, fast and slow responses would need to be distinguished on top of distinguishing problem-skipping from ability. Thus, the current IRTree model should be extended to include a node which encompasses fast from slow response processes [7].

Similarly, many studies aimed at separating item omissions from ability in traditional educational contexts also distinguish between items that are skipped and items that are not reached ([14, 31, 11]). Not-reached items are typically defined as items that the student did not get to in time before the deadline of the test. In our data, such items cannot be defined because whole sessions do not have a time limit. However, they can be incorporated into the model in two ways. First, individual items which are not responded to in time (i.e., the user has a time-out response) can be modeled separately from ability. This would allow us to test whether the same mechanism underlies ability and responding in time. Second, in Math Garden, a student can choose to end a session before they have finished all 10 items. While the adaptive system should ensure that all students play at a level that matches their ability, the IRTree model can be extended to check whether users who quit a session prematurely have different ability estimates compared to users who finish a full session. Huang [19] used an extension of an IRTree model to distinguish aberrant from normal test-taking behavior, where aberrant behavior was classified by dropping out, skipping items, or showing a gradual decline in exerted effort. Similarly, such a model can be applied to online learning to detect differences in motivational states between students. Altogether, the IRTree framework has proven to be a useful method in estimating the latent correlation between two different response types in online learning data, with interesting options for extensions.

Finally, it is important to note 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. Future work would benefit from comparing different modeling frameworks aimed at deciphering the response process of students in an educational context. For example, Race Models [15] 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 [15, 12]. Alternative methods to model students' decision-making processes include Bayesian Hierarchical Models [37, 25] and Markov Decision Processes [24, 32]. 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, given that most educational measurement models use IRT, IRTrees are an ideal candidate as they can be directly compared and offer estimates that are easy to interpret in, e.g., teacher dashboards.

6 Conclusion

In a large-scale adaptive learning environment for Arithmetic practice, low-ability students were more likely to skip problems. The measurement model accounting for this relationship performed best in estimating student ability. Under a measurement model that assumes that problem-skipping and accuracy are unrelated, student ability may be overestimated, leading adaptive systems to assign overly difficult items. This misalignment increases the likelihood of further problem-skipping, reinforcing the bias in ability estimation. To prevent this, problem-skipping should be measured separately from accuracy. Online learning environments that utilize problem-skipping options 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 problem-skipping is reduced. Only with accurate measurement models can equity in learning analytics be realized.

Acknowledgements

Annie M. Johansson is supported by the Ministry of Education, Culture and Science (Netherlands).

Alexander O. Savi received funding from the Dutch Research Council (VI.Veni.211G.016).

References

[1] Ryan Shaun Baker, Albert T Corbett, Kenneth R Koedinger, and Angela Z Wagner. Off-task behavior in the cognitive tutor classroom: When students "game the system". In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 383–390, 2004. doi: https://doi.org/10.1145/985692.985741.

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

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

[4] Ulf Böckenholt and Thorsten Meiser. Response style analysis with threshold and multi-process irt models: A review and tutorial. *British journal of mathematical and statistical psychology*, 70(1):159–181, 2017. doi: https://doi.org/10.1111/bmsp.12086.

[5] Matthieu Brinkhuis, Wessel Cordes, and Abe Hofman. Governing games: Adaptive game selection in the math garden. In *ITM Web of Conferences*, volume 33, page 03003. EDP Sciences, 2020. doi: https://doi.org/10.1051/itmconf/20203303003.

[6] Matthieu JS Brinkhuis, Alexander O Savi, Abe D Hofman, Frederik Coomans, Han LJ van Der Maas, and Gunter Maris. Learning as it happens: A decade of analyzing and shaping a large-scale online learning system. *Journal of Learning Analytics*, 5(2):29–46, 2018. doi: http://dx.doi.org/10.18608/jla.2018.52.3.

[7] Frederik Coomans, Abe Hofman, Matthieu Brinkhuis, Han LJ van der Maas, and Gunter Maris. Distinguishing fast and slow processes in accuracy-response time data. *PloS one*, 11 (5):e0155149, 2016. doi: https://doi.org/10.1371/journal.pone.0155149.

[8] Paul De Boeck and Ivailo Partchev. Irtrees: Tree-based item response models of the glmm family. *Journal of Statistical Software*, 48:1–28, 2012. doi: https://doi.org/10.18637/jss.v048.c01.

[9] Susanne MM de Mooij, Maartje EJ Raijmakers, Iroise Dumontheil, Natasha Z Kirkham, and Han LJ van der Maas. Error detection through mouse movement in an online adaptive learning environment. *Journal of Computer Assisted Learning*, 37(1):242–252, 2021. doi: https://doi.org/10.1111/jcal.12483.

[10] Susanne MM de Mooij, Iroise Dumontheil, Natasha Z Kirkham, Maartje EJ Raijmakers, and Han LJ Van Der Maas. Post-error slowing: Large scale study in an online learning environment for practising mathematics and language. *Developmental science*, 25(2):e13174, 2022. doi: https://doi.org/10.1111/desc.13174.

[11] Dries Debeer, Rianne Janssen, and Paul De Boeck. Modeling skipped and not-reached items using irtrees. *Journal of Educational Measurement*, 54(3):333–363, 2017. doi: https://doi.org/10.1111/jedm.12147.

[12] Nathan J Evans, Mark Steyvers, and Scott D Brown. 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, 2018. doi: https://doi.org/10.1111/cogs.12627.

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

[14] Cees A. W. Glas and Jonald L. Pimentel. Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement*, 68(6):907–922, 2008. doi: http://journals.sagepub.com/doi/10.1177/0013164408315262.

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

[16] Anna Hilz, Karin Guill, Janina Roloff, Karen Aldrup, and Olaf Köller. The relationship between individual characteristics and practice behaviour within an adaptive arithmetic learning program. *Journal of Computer Assisted Learning*, 39(3):970–983, 2023. doi: https://doi.org/10.1111/jcal.12780.

[17] Abe D Hofman, Brenda RJ Jansen, Susanne MM De Mooij, Claire E Stevenson, and Han LJ Van der Maas. A solution to the measurement problem in the idiographic approach using computer adaptive practicing. *Journal of Intelligence*, 6(1):14, 2018. doi: https://doi.org/10.3390/jintelligence6010014.

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

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

[20] Brenda RJ Jansen, Jolien Louwerse, Marthe Straatemeier, Sanne HG Van der Ven, Sharon Klinkenberg, and Han LJ Van der Maas. The influence of experiencing success in math on math anxiety, perceived math competence, and math performance. *Learning and individual differences*, 24:190–197, 2013. doi: https://doi.org/10.1016/j.lindif.2012.12.014.

[21] Brenda RJ Jansen, Eva A Schmitz, and Han LJ Van der Maas. Affective and motivational factors mediate the relation between math skills and use of math in everyday life. *Frontiers in psychology*, 7:513, 2016. doi: https://doi.org/10.3389/fpsyg.2016.00513.

[22] Rogier A Kievit, Abe D Hofman, and Kate Nation. Mutualistic coupling between vocabulary and reasoning in young children: A replication and extension of the study by kievit et al.(2017). *Psychological science*, 30(8):1245–1252, 2019. doi: https://doi.org/10.1177/0956797619841265.

[23] Sharon Klinkenberg, Marthe Straatemeier, and Han LJ van der Maas. Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, 57(2):1813–1824, 2011. doi: https://doi.org/10.1016/j.compedu.2011.02.003.

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

[25] Michael D Lee and Eric-Jan Wagenmakers. *Bayesian cognitive modeling: A practical course*. Cambridge university press, 2014.

[26] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

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

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

[29] Taichi Okumura. 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, 2014. doi: https://doi.org/10.1177/0013164413516976.

[30] Lu Ou, Abe D Hofman, Vanessa R Simmering, Timo Bechger, Gunter Maris, and Han LJ van der Maas. Modeling person-specific development of math skills in sontinuous time: New evidence for mutualism. *International Educational Data Mining Society*, 2019.

[31] Steffi Pohl, Linda Gräfe, and Norman Rose. 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, 2014. doi: https://doi.org/10.1177/0013164413504926.

[32] Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994.

[33] Georg Rasch. *Probabilistic models for some intelligence and attainment tests.* ERIC, 1993.

[34] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi: https://doi.org/10.1093/biomet/63.3.581.

[35] Alexander O Savi, Nienke M Ruijs, Gunter KJ Maris, and Han LJ van der Maas. 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, 2018. doi: https://doi.org/10.1016/j.compedu.2017.12.008.

[36] Alexander O Savi, Chris van Klaveren, and Ilja Cornelisz. Combating effort avoidance in computer adaptive practicing: Does a problem-skipping restriction promote learning? *Computers & Education*, 206:104908, 2023. doi: https://doi.org/10.1016/j.compedu.2023.104908.

[37] Robert Sawyer, Jonathan Rowe, Roger Azevedo, and James Lester. Modeling player engagement with bayesian hierarchical models. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 14, pages 257–263, 2018. doi: https://doi.org/10.1609/aiide.v14i1.13048.

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

[39] Hanke Vermeiren, Abe D Hofman, Maria Bolsinova, Han LJ van der Maas, and Wim Van Den Noortgate. Balancing stability and flexibility: Investigating a dynamic k value approach for the elo rating system in adaptive learning environments. *OSF*, 2024. doi: https://doi.org/10.31234/osf.io/6hzke. Preprint.

[40] Dan Ye. The history and development of learning analytics in learning, design, & technology field. *TechTrends*, 66(4):607–615, 2022. doi: https://doi.org/10.1007/s11528-022-00720-1.

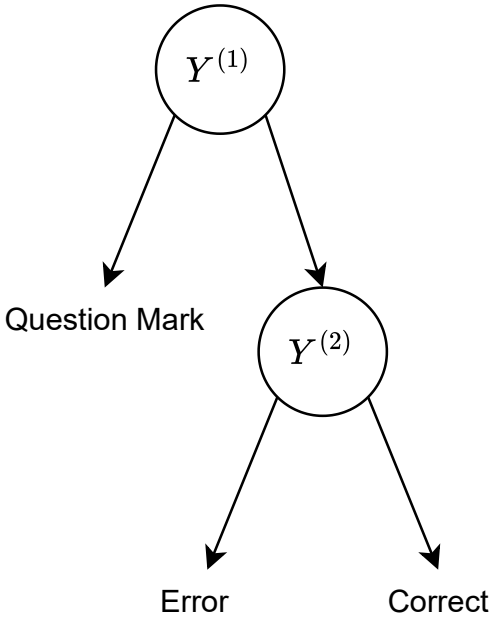

Figure 1: A binary decision tree (IRTree model) with 2 nodes reflecting the decision to respond or not (Y_1) and whether the response is correct or not (Y_2).

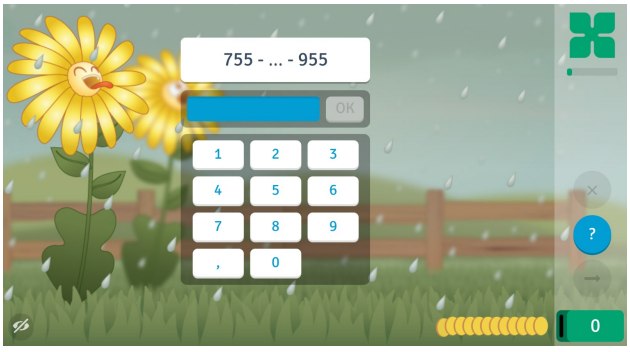

Figure 2: A screenshot of the Math Garden game Series. A student is asked to fill in the missing number in the sequence. They gain (correct answer) or lose (incorrect answer) coins corresponding to the remaining amount of time in seconds (bottom right). A student can click the question mark (middle right)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

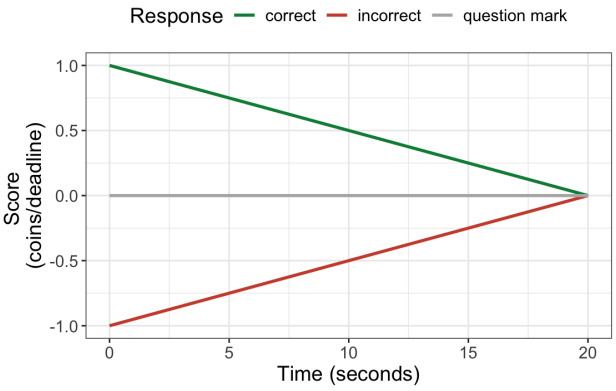

Figure 3: Visualization of the explicit scoring rule in Math Garden.

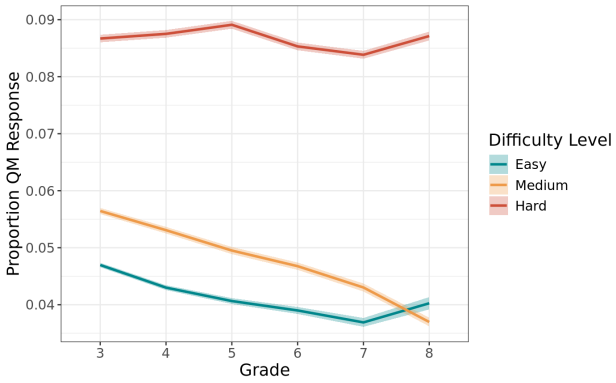

Figure 4: Proportion of question mark responses in the Series game, separated by grade and difficulty level (easy = 90% correct; medium = 75% correct; hard = 60% correct). Ribbons around the line denote the 95% confidence interval. Grade refers to the Dutch grade levels; age ranges approximately between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

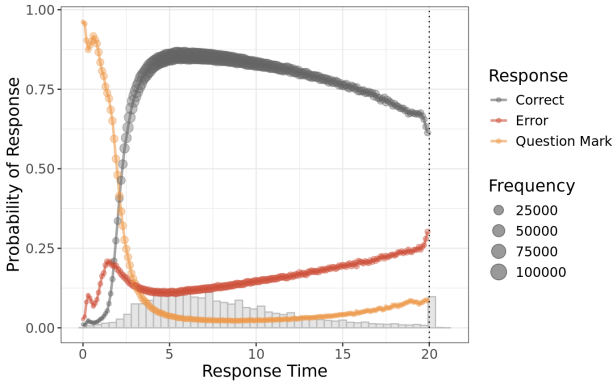

Figure 5: The probability of each response type (correct, error, question mark) across reponse times (seconds). Point size refers to the amount of observations at that data point. The vertical line denotes the deadline to respond, at 20 seconds. The histogram displays the average distribution of response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

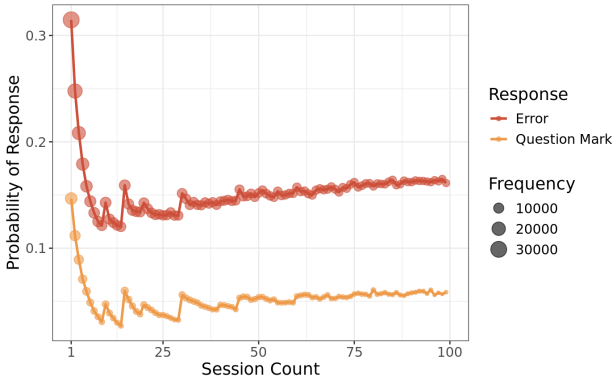

Figure 6: The probability of answering with a question mark or giving an incorrect response, over the first 100 items played in the Series game. Point size refers to the amount of observations at that data point.

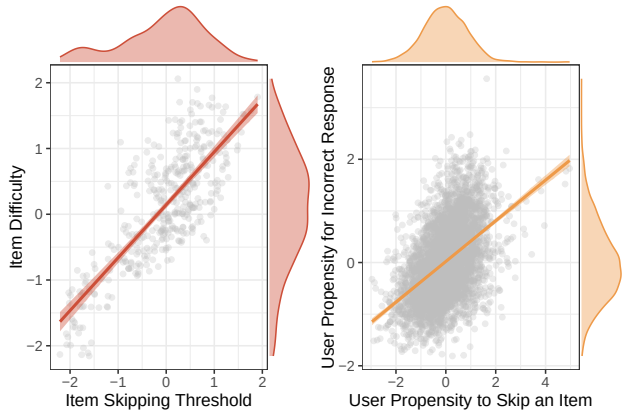

Figure 7: Correlation between estimated nodes in the fully estimated IRTree model. The left graph displays the correlation between the item skipping threshold ($\beta^{(1)}$), and the item difficulty ($\beta^{(2)}$). The right-hand graph displays the correlation between the individual propensity to skip 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.

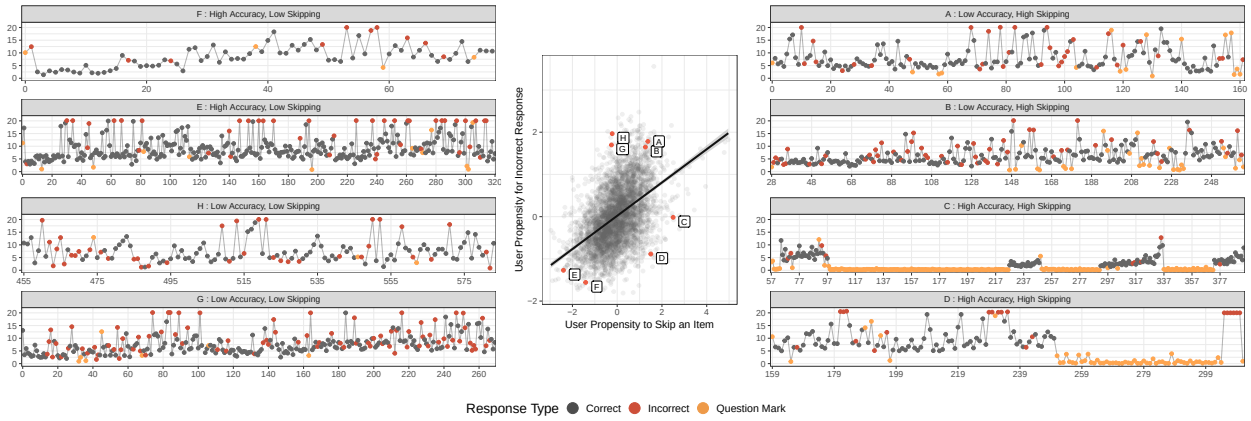

Figure 8: This graph displays eight individual users extracted based on their estimated propensity to skip and make an incorrect response (center graph). Users are displayed in a clockwise fashion starting from the upper right corner (low accuracy, high skipping), to the upper left corner (low accuracy, 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.