Measurement of Ability in Adaptive Learning and Assessment Systems when Learners Use On-DemandHints

Maria Bolsinova1, Benjamin Deonovic2, Meirav Arieli-Attali3, Burr Settles4, Masato Hagiwara5, and Gunter Maris6

Abstract
Adaptive learning and assessment systems support learners in acquiring knowledge and skills in a particular domain. The learners’ progress is monitored through them solving items matching their level and aiming at specific learning goals. Scaffolding and providing learners with hints are powerful tools in helping the learning process. One way of introducing hints is to make hint use the choice of the student. When the learner is certain of their response, they answer without hints, but if the learner is not certain or does not know how to approach the item they can request a hint. We develop measurement models for applications where such on-demand hints are available. Such models take into account that hint use may be informative of ability, but at the same time may be influenced by other individual characteristics. Two modeling strategies are considered: (1) The measurement model is based on a scoring rule for ability which includes both response accuracy and hint use. (2) The choice to use hints and response accuracy conditional on this choice are modeled jointly using Item Response Tree models. The properties of different models and their implications are discussed. An application to data from Duolingo, an adaptive language learning system, is presented. Here, the best model is the scoring-rule-based model with full credit for correct responses without hints, partial credit for correct responses with hints, and no credit for all incorrect responses. The second dimension in the model accounts for the individual differences in the tendency to use hints.

1Tilburg University, Tilburg, Netherlands
2Corteva Agriscience, Johnston, IA, USA
3Center for Educational Technology, Tel Aviv, Israel
4Duolingo, Pittsburg, PA, USA
5Octanove Labs LLC, Seattle, WA, USA
6Tata Consultancy Services, Zaventem, Belgium

Corresponding Author:
Maria Bolsinova, Tilburg University, Department of Methodology and Statistics, Prof. Cobbenhagenlaan 225, 5037 DB Tilburg, the Netherlands.
Email: bolsinova@gmail.com
Keywords
adaptive learning systems, item response tree models, measurement, multidimensional nominal response model, hints, item response theory, measurement theory

Introduction

Adaptive learning and assessment systems are designed to dynamically adjust the level or type of practice and instruction material based on an individual learner’s abilities or skills and other characteristics. In such systems, accurate and precise measurement of the learners’ level of ability is important since monitoring the development of learners’ skills is crucial to adapt the learning material to their level. The level of learners’ ability is usually estimated based on the correctness of the responses that they give to the practice items. However, information on learners’ behavior in the system, and in particular whether learners use hints, can be used to obtain additional information about the measured abilities or skills, and can be included in the scoring rule when estimating ability. Hint use can be considered as process data as opposed to the response accuracy which constitutes product data from the assessment. While a large number of models have been developed to incorporate different types of process data into measurement, for example response times (for an overview see e.g., Kyllonen & Zu, 2016), measurement models for applications in which respondents may request hints instead of directly answering items could be more fully developed and explored.

Previous research on modeling hint use of learners yield different approaches. One of the early works by Aleven and colleagues (e.g., Aleven & Koedinger, 2000; Aleven et al., 2004) took the approach of predefining productive rules for help-seeking behavior (i.e., the ideal behavior) and examined whether learners adhere to these rules. This normative model of help-seeking suggests that there are appropriate times for requesting hints. In some conditions using hints including bottom-out-hints that show the correct response is warranted, particularly when students are at a low level of skill acquisition. The authors found that the majority of learners (72%) did not follow these productive rules, but rather exhibit what they called “unproductive behavior” and “hint abuse,” where learners used the hints to get at the answer rather than to learn. It may be noted, however, that using bottom-out hints does not necessarily have to be considered unproductive, as they can also provide new learners with worked examples of how to solve problems (see Shih et al., 2011, for a model that distinguishes productive use of bottom-out hints from hint abuse by means of response times).

A different approach was taken by Feng and Heffernan (2010), Wang and Heffernan (2011), and Wang et al. (2010). In their research, a data-driven approach was taken without predetermined assumptions about hint use, except to assume that the more hints used by the learner, the lower the probability that they possess the knowledge. The model they proposed—the Assistance Model (AM; Wang & Heffernan, 2011)—uses the information about the number of hints used and the number of answer attempts till correct to predict the performance of learners on the next item. Since the AM seemed to predict the performance on the next item, the researchers added the model to other models that track learning, combined with a Bayesian Knowledge Tracing model (Wang & Heffernan, 2011) and a partial credit model (Wang et al., 2010).

When modeling hint use behavior, Goldin et al. (2013) applied a multinomial logistic regression that distinguishes hint-request tendency from proficiency. They found that seeking just one hint is associated with repeated hint-seeking, but when students do make attempts to solve a problem after viewing a hint, they succeed about half of the time. They also found that the individual differences in the tendency to use hints are stronger than the differences in proficiency.
The approaches to modeling hint use discussed so far focus on the prediction of the correctness of the next response. Another perspective that has been taken by different authors (e.g., Cen et al., 2006; Ueno & Miyazawa, 2017) and is considered in this paper focuses on the measurement of ability in the presence of hints. For measurement of ability, item response theory (IRT) models are often used (for an overview see e.g., Hambleton & Swaminathan, 2013; van der Linden & Hambleton, 2013). These are statistical models that relate the probability of a correct response to an item to the persons’ ability, which allows the estimation of the level of ability given the observed responses. One of the most common IRT models is the two-parameter logistic model (2PL; Birnbaum, 1968) in which the probability of a correct response to an item is modeled as a function of the person’s ability and the item characteristics:

$$
Pr(X_i = 1|\theta) = \frac{\exp(\alpha_i(\theta - \delta_i))}{1 + \exp(\alpha_i(\theta - \delta_i))},
$$

(1)

where $X_i$ is the binary response accuracy on item $i$ (1—correct, 0—incorrect), $\theta$ is the ability latent variable, $\delta_i$ is the difficulty of the item, and $\alpha_i$ is the discrimination of the item. The response accuracies on different items are assumed to be conditionally independent given the latent variable. The density of the data from $N$ learners to $K$ items is the following:

$$
f(a, \delta; x, z) = \prod_p \int \prod_i \left( \frac{\exp(x_{pi}a_i(\theta - \delta_i))}{1 + \exp(a_i(\theta - \delta_i))} \right)^{z_{pi}} N(\theta; 0, 1) d\theta,
$$

(2)

where $x$ is the partially complete data matrix of response accuracy with $x_{pi}$ being the response accuracy of person $p$ to item $i$, $z$ is the $N \times K$ matrix with each element $z_{pi}$ indicating whether learner $p$ responded to item $i$ (1—yes, 0—no), $a$ and $\delta$ are the vectors of length $K$ of the item discrimination and difficulty parameters, and the ability is assumed to have a normal distribution in the population of learners with the mean and variance constrained to 0 and 1 for identification purposes.

When in an adaptive learning environment hints are provided to learners, the standard measurement model needs to be adapted to include the hint use. When adapting the measurement model, it is important to distinguish the two ways in which hints can be presented to learners: (1) providing hints after an incorrect response, or (2) providing hints on-demand. These two ways of providing hints differ both from the learning perspective and with respect to how hints should be incorporated in the measurement model. The differential learning effect of providing hints after an incorrect response or providing hints on-demand has been studied by Razzaq and Heffernan (2010). They found that the students learned reliably more with hints on-demand than with hints after an incorrect response. On-demand hints might also be preferable as a tool for improving the learners’ motivation by providing them with more control of their learning process and giving them more freedom within the learning environment which might be advantageous for the learning process.

The different ways of providing hints also require different statistical treatment. Let us by $y_{pi}$ denote a binary indicator of whether the learner $p$ was provided with a hint when responding to item $i$. When a hint is provided after an incorrect response, the hint use itself is not a random variable since $y_{pi} = 1 - x_{pi1}$ where $x_{pi1}$ is the response accuracy on the first attempt to answer item $i$. That is, if the response is correct on the first attempt the hint is never provided, while if the response on the first attempt is incorrect then the hint is always provided. To include this in a measurement model, one can instead of modeling the distribution of the binary response accuracy as in the 2PL model consider the distribution of the accuracy variable with three levels: $X_i^* = 2$ if the response is correct after the first attempt, $X_i^* = 1$ if the response is incorrect after the first attempt, but correct after the hint was provided in the second attempt, and $X_i^* = 0$ if the response is
incorrect in both attempts. The distribution of $X^*_i$ given the ability latent variable can be modeled using IRT models for polytomous data, for example the partial credit model (Masters, 1982) or the graded response model (Samejima, 1969), as has been done by Ueno and Miyazawa (2017).

When hints are provided on demand, the hint use $Y_i$ on item $i$ is itself a random variable, since the learner is always free to choose whether to request a hint or not. In this case, instead of having one variable $X^*_i$ with three possible outcomes, for each item there are two binary random variables ($X_i$ and $Y_i$) with four possible outcomes in total. One approach for dealing with this is to treat the use of hints similar to incorrect responses. Then, a variable $X^{**}_i = I(X_i = 1)I(Y_i = 0)$ (i.e., 1 if the response is correct without a hint, and 0 otherwise) can be modeled using dichotomous IRT models or their generalizations like the Additive Factors Model (Cen et al., 2006). An alternative, which we are going to pursue in this paper, is to model the full joint distribution of $X_i$ and $Y_i$ without collapsing the different types of outcomes. Our motivation for this is that not only the accuracy of the responses but also whether students use hints may be informative about ability. At the same time, there might be other individual differences between learners with respect to hint use behavior that are not directly related to ability, which should also be taken into account when modeling the joint distribution of response accuracy and hint use. The goal of this paper is to develop measurement models for ability for applications where hints are provided on-demand that allow one to estimate learner’s level of ability based on both response accuracy and hint use behavior but at the same time acknowledge that there might be other relevant differences between the learners influencing their interactions with the learning system.

The rest of the paper is organized as follows. The Data section describes the adaptive language learning system Duolingo, data from which is used to illustrate the different approaches to the joint modeling of response accuracy and hint use. The details on the dataset used in the study are provided. In the section Jointly modeling hint use and response accuracy, we propose different modeling approaches for measurement of ability when hints are available on-demand. The Methods section describes the data analysis methods. The Results and Discussion section presents and discusses the results of applying different measurement models to the data from Duolingo. The paper ends with general conclusions.

Data

Data with hint use was obtained from Duolingo, an online language learning platform with more than 200 million registered users. A Duolingo course is organized into a tree-like structure called a skill tree. A node in this tree is called a skill. A row in a course is a set of skills which have the same depth in the skill tree. Users can move on to the next row only after they complete all skills in the previous row. After starting a lesson, users learn the language by solving a series of exercises (i.e., items) called challenges. A consecutive series of challenges is called a session. Users can also choose to practice the skill they already completed, by pressing a practice button on a skill page. This is called a skill practice session, which shows one a series of challenges on the words and grammar that need reviewing the most according to Duolingo’s internal word strength model.

The dataset from Duolingo available for analysis contained items completed by learners between November 9, 2015, and December 8, 2015. Only users who created a new account during the data collection period and reached the tenth row in their respective course’s skill tree were included in the sample. From the six language courses that were available for analysis, two (with the largest number of learners) were included in this study. One dataset (from the course for learning Spanish for English speakers, i.e., Spanish-from-English) was used for preliminary analysis to get insights about what the important characteristics of the data are which should be taken into account when developing the models for hint use and response accuracy. Another
dataset (English-from-Portuguese) was used after the models were developed to select the model best describing the data.

For data analysis subsets of the data were selected. First, for each of the two courses only data from a single platform (with the largest number of users) was selected to avoid the additional differences between the learners from the different user interfaces on different platforms: data from Android was used in the Spanish-from-English course, and data from iOS was used for the English-from-Portuguese course. Second, only challenges which were the translate challenge (from foreign language to native language) were kept and only the first time a user responded to this challenge was used. These translation items give the learners the opportunity to ask for a hint by checking the translation of one or more words in the sentence. The learner may hover over any word in the sentence and see the translation to their native language. If a learner hovers over at least one of the words, this is registered as a learner using a hint on an item \( Y_{pi} = 1 \). The learner has to type their response and the accuracy of the response is recorded. The item is considered correct if the sentence is translated correctly.\(^1\)

The subset of users and items included 2639 learners and 8848 items for Spanish-from-English, and 5264 learners and 7087 items for English-from-Portuguese. Some additional selection criteria were applied to the learners and the items. Only the items which are complete sentences (i.e., not word combinations or separate words) and have more than two non-article words were included. Items with less than 40% observed responses were removed. For the remaining items, users who had more than 10 responses were selected. Finally, to make sure that there is not a lot of overlap in words between the selected sentences, for each item pair for which there was more than 70% overlap in words the shortest item was removed. This resulted in 89 items and 1109 users for Spanish-from-English and 99 items and 3845 learners for English-from-Portuguese.

**Jointly modeling hint use and response accuracy**

*Initial model: Hint use as part of the scoring rule*

Information on whether learners use hints can be considered process data as opposed to the response accuracy which constitutes product data from the assessment. Process data can be used to obtain additional information about the measured abilities or skills, and can be included in the scoring rule when estimating ability. This has been done in the context of response times, which is another example of process data. Maris and van der Maas (2012) proposed a model for jointly modeling response accuracy and response time called signed-residual-time model in which the contribution of response time in the item score depends on the accuracy of the response:

\[
S = \sum_i (2X_i - 1)(d - T_i),
\]

where \( S \) is the total score of a person, \( X_i \) is response accuracy on item \( i \), \( T_i \) is response time, and \( d \) is the time limit. The remaining time before the deadline is either added to the score if the response is correct, or subtracted from the score if the response is incorrect. In this way, the scoring rule encourages fast correct responses, but discourages fast incorrect responses. An IRT model has been developed based on this scoring rule and the assumption of conditional independence between the scores on different items given ability, which allows to estimate the differences between the items in their difficulty and the differences between the persons in their ability based both on their response accuracy and response time.

The same idea can be used in the context of hint use. The contribution of the item to the score can be based on both whether the response was correct and on whether it was obtained with a hint.
With two binary variables ($X_i$ and $Y_i$), four different outcomes can be considered, each matching an item score:

$$S_{pi} = \begin{cases} 0 & \text{if } X_{pi} = 0, Y_{pi} = 0; \\ 1 & \text{if } X_{pi} = 0, Y_{pi} = 1; \\ 2 & \text{if } X_{pi} = 1, Y_{pi} = 1; \\ 3 & \text{if } X_{pi} = 1, Y_{pi} = 0. \end{cases}$$

(4)

Higher scores (2 and 3) are given to correct responses, and the impact of hint use on the item score depends on whether the response is correct (similarly to the differential contribution of response time in the signed-residual-time model): If the response is correct, then a higher score is obtained when no hints were used, while if the response is incorrect, then a higher score is obtained when a hint was used. Similarly to how response time is treated in the signed-residual-time model, here correct responses without hints are encouraged, while incorrect responses without hints are discouraged.

An IRT model can be derived from this scoring rule for which the total score for the person $\sum S_{pi}$ is a sufficient statistic for the person’s ability, and $\sum S_{pi}$ is the sufficient statistic for the item’s difficulty (see Appendix for the details of the derivation):

$$\Pr(S_i = s | \theta) = \frac{\exp(s(\theta - \delta_i))}{\sum_{r=0}^{3} \exp(r(\theta - \delta_i))},$$

(5)

where $s \in \{0, 1, 2, 3\}$, $\theta$ is ability latent variable, and $\delta_i$ is the difficulty of item $i$. This is a constrained version of the partial credit model in which there is a single item difficulty parameter instead of multiple threshold parameters. This single parameter can be interpreted as follows: When ability of a person is equal to its value then all four outcomes are equally likely. The higher $\delta_i$ is, the higher the expected score on the item is.

**Extending the model**

The model in Equation 5 implies that the items do not differ in their discriminatory power, that is the strength of the relationship between the item score and ability is assumed to be the same across items. However, it might be beneficial to extend the model for hint use and accuracy to allow for the items to differ in the strength of the relationship between the item score and ability, similarly to how the Rasch model (Rasch, 1960) has been extended to the 2PL model and the signed-residual-time model has been extended to its generalized version (van Rijn & Ali, 2018). The extended model is:

$$\Pr(S_i = s | \theta) = \frac{\exp(s\alpha_i(\theta - \delta_i))}{\sum_{r=0}^{3} \exp(r\alpha_i(\theta - \delta_i))},$$

(6)

where $\alpha_i > 0$ is the discrimination parameter of item $i$.

The model in Equation 6 implies the following conditional distributions of response accuracy given hint use:

$$\Pr(X_i = 1 | \theta, Y_i = 1) = \Pr(S_i = 2 | \theta, S_i \in \{1, 2\}) = \frac{\exp(\alpha_i(\theta - \delta_i))}{1 + \exp(\alpha_i(\theta - \delta_i))}$$

(7)

for when the hints were used, and
for when the hints were not used. In both cases, it is a 2PL model with the same difficulty ($\delta_i$), but different discrimination ($\alpha_i$ and $3\alpha_i$, respectively). Note, that the difficulty of item $i$ can be interpreted as such a point on ability scale for which the probabilities of obtaining the four different scores on the item are all equal to $1/4$. Thus, the model poses a constraint on the conditional difficulties of the items. To evaluate whether such a constraint makes sense in the context of Duolingo data, we fitted the 2PL model separately to the Spanish-from-English accuracy data of those responses where hints were used, and those responses where hints were not used. Figure 1 shows the estimated conditional difficulties of the items. Since the dots do not lie on the identity line, we conclude that it might be beneficial to extend the model to allow for the conditional accuracy models to have separate difficulty parameters. The model can be extended as follows:

$$
\Pr(X_i = 1|\theta, Y_i = 0) = \Pr(S_i = 3|\theta, S_i \in \{0, 3\}) = \frac{\exp(3\alpha_i(\theta - \delta_i))}{1 + \exp(3\alpha_i(\theta - \delta_i))}
$$

Figure 1. Item difficulties in the Duolingo Spanish-from-English data set estimated separately from the responses in which no hints were used (on the $x$-axis) and from the responses in which hints were used (on the $y$-axis). The parameters were estimated for the two-parameter logistic model using marginal maximum likelihood and the R-package mirt (Chalmers, 2012). The estimates do not lie on the identity line, which means that it is not reasonable to assume that the difficulty of the items is independent on hint use.

$$
\Pr(S_{pi} = s|\theta) = \frac{\exp(s\alpha_i\theta + \delta_{is})}{\sum_{r=0}^{3}\exp(r\alpha_i\theta + \delta_{ir})},
$$

where $\delta_{is}$ is a category-specific parameter with $\delta_{i0} = 0$ for identification purposes. With this extension the model for the item score is the generalized partial credit model Muraki (1992). Here, the conditional difficulty given no hint use is equal to ($\delta_{i1} - \delta_{i2}$), and the conditional difficulty given hint use is $-\delta_{i3}/3$.

The model in Equation 9 implies positive correlations between the item scores from different items ($S_i$ and $S_j$, $\forall i, j$) because all of them are positively correlated with the latent ability. It also implies positive correlations between the response accuracies on different items ($X_i$ and $X_j$, $\forall i, j$). However, it does not necessarily imply positive correlations between the hint use variables on different items ($Y_i$ and $Y_j$, $\forall i, j$). Hint use variables on items $i$ and $j$ are positively correlated under the model when the response accuracy on these items is the same, but negatively correlated if one
item response was correct and the other was incorrect. This relationship is derived from the fact that given a correct response, \( Y_i \) is negatively related to the latent ability \( \theta \):

\[
\Pr(Y_i = 1|\theta, X_i = 1) = \Pr(S_i = 2|\theta, S_i \in \{2, 3\}) = \frac{\exp(-\alpha_i \theta + (\delta_{i2} - \delta_{i3}))}{1 + \exp(-\alpha_i \theta + (\delta_{i2} - \delta_{i3}))};
\]

(10)

while given an incorrect response, \( Y_i \) is positively related to \( \theta \):

\[
\Pr(Y_i = 1|\theta, X_i = 0) = \Pr(S_i = 1|\theta, S_i \in \{0, 1\}) = \frac{\exp(\alpha_i \theta + \delta_{i1})}{1 + \exp(\alpha_i \theta + \delta_{i1})}.
\]

(11)

While the model does not generally imply positive correlations between \( Y_i \) and \( Y_j \), the tetrachoric correlations between hint use variables in the Spanish-from-English Duolingo data set are generally positive, see the histogram of these correlations in Figure 2. Therefore, it might be beneficial to extend the model to allow for these kind of correlations between the hint use variables. To account for these positive correlations, we extend the model with an additional latent variable related to the individual differences in hint use: Learners with high levels of this latent variable tend to use hints on most of the items, while learners with low levels of this latent variable tend to respond to the items without using hints. The multidimensional nominal response model (Takane & De Leeuw, 1987; Thissen & Cai, 2016) can be used to model the response accuracy and hint use where these two variables relate to the measured ability through the same scoring rule as in the model in Equation 9, and additionally the hint use variable is related to the additional latent variable:

\[
\Pr(S_i = s|\theta, \eta) = \frac{\exp(s\alpha_i \theta + \mathcal{I}(s \in \{1, 2\}) \lambda_i \eta + \delta_{is})}{\sum_{r=0}^{3} \exp(r\alpha_i \theta + \mathcal{I}(r \in \{1, 2\}) \lambda_i \eta + \delta_{ir})},
\]

(12)

where \( \eta \) is the extra latent variable accounting for the differences in hint use, \( \lambda_i > 0 \) is the item loading for this latent variable, and \( \mathcal{I}(\text{condition}) \) is the identity function which takes a value of one if the condition is satisfied, and a value of zero if it is not. Here the conditional distributions of accuracy are the same as in the model in Equation 9, but the conditional distributions of hint use depend on both latent variables, ability and tendency to use hints (\( \eta \)).

Figure 2. Histogram of the tetrachoric correlations between hint use variables on different items in the Spanish-from-English Duolingo data set. Most of the correlations are positive and some of them are relatively high, which means that it might be beneficial to extend the scoring-rule-based model with an additional dimension which would account for these correlations.
The model in Equation 12 can be alternatively formulated as follows:

\[
\Pr(S_i = s | \theta, \eta) = \frac{\exp(a_s \alpha_i \theta + b_s \lambda_i \eta + \delta_i s)}{\sum_{r=0}^{3} \exp(a_r \alpha_i \theta + b_r \lambda_i \eta + \delta_i r)}.
\] (15)

where \(a = [0, 1, 2, 3]\) is the scoring function for the ability latent variable, and \(b = [0, 1, 1, 0]\) is the scoring function for the tendency to use hints.\[\]

**Alternative scoring rules**

In the previous two subsections we have taken as the starting point that the appropriate scoring rule for the ability is to assign the most points to the correct response without hints, followed by the correct response with hints, the incorrect response with hints, and finally the lowest credit is given to incorrect responses without hints. This choice was made because we used the signed-residual-time model for response time and accuracy as a source of inspiration. We hypothesized that hint use is analogous to the use of extra time in the sense that it can be considered a resource that learners can use, and it should therefore be incorporated in the scoring rule in a similar way. However, one could argue that scores for different outcomes should be assigned in a different way. Here, we consider four alternative scoring rules:

1. Incorrect responses without hints are better than incorrect responses with hints.
2. Correct responses without hints are better than correct responses with hints, but for incorrect responses there is no difference in the scoring functions for ability between the responses with and without hints.
3. Incorrect responses without hints are better than the responses with hints. The idea here is that learners with the lowest ability will not attempt to answer the items without hints, learners with a bit higher levels of ability will attempt to answer without hints, but will often fail, and the learners with high levels of ability will attempt to answer without hints and will often succeed.
4. Only correct responses without hints receive full credit, while all other options (i.e., either an incorrect response, or using a hint) receive no credit.

Each of the alternative scoring rules can be translated in the multidimensional nominal response model with different specifications for \(a\). For the three alternative scoring rules, \(a\) is equal to \([1, 0, 2, 3]\), \([0, 0, 1, 2]\), \([2, 0, 1, 3]\), and \([0, 0, 0, 1]\), respectively, and \(b = [0, 1, 1, 0]\) in all of the models. The last scoring rule corresponds to the common way of modeling responses in the presence of hints (e.g., Cen et al., 2006).

**IRTTree modeling approach as an alternative to the scoring-rule-based approach**

So far, we considered a scoring-rule-based approach to modeling the joint distribution of response accuracy and hint use. Alternatively, when modeling this joint distribution one might consider the observations as outcomes of two sequential events: 1. A learner decides whether to use a hint or
not; 2. A learner gives a response (depending on the outcome of the first event it is a response based on the hint or no hint), which can be either correct or incorrect. Here, the joint distribution of \( X_i \) and \( Y_i \) can be factorized as the product of the marginal distribution of \( Y_i \) (i.e., modeling the outcome of the first event) and the conditional distribution of \( X_i \) given \( Y_i \) (i.e., modeling the outcome of the second event conditional on the outcome of the first one). This can be presented in a tree-structure (see Figure 3). Following the IRTree modeling approach of Boeck and Partchev (2012) and the generalized IRTree model of Jeon and De Boeck (2016), each node of the tree can be modeled using an IRT model. Particularly, one can model the first node of the IRTree with a 2PL model in which \( Y_i \) depends on the latent variable quantifying the learners’ tendency to use hints (\( \eta \)):

\[
\Pr(Y_i = 1 | \eta) = \frac{\exp(a_{i1} \eta + \beta_{i1})}{1 + \exp(a_{i1} \eta + \beta_{i1})},
\]

where the intercept \( \beta_{i1} \) allows to model the differences between the items in terms of how often hints are used, and the slope \( a_{i1} \) quantifies the differences between the items in the strength of the relationship between the probability of a hint being used and the learners’ tendency to use hints. Analogously, for the second and the third node of the IRTree a 2PL model can also be used:

\[
\Pr(X_i | Y_i = 1, \theta_1) = \frac{\exp(a_{i2} \theta_1 + \beta_{i2})}{1 + \exp(a_{i2} \theta_1 + \beta_{i2})},
\]

\[
\Pr(X_i | Y_i = 0, \theta_2) = \frac{\exp(a_{i3} \theta_2 + \beta_{i3})}{1 + \exp(a_{i3} \theta_2 + \beta_{i3})},
\]

where \( \theta_1 \) and \( \theta_2 \) are potentially different latent variables related to the accuracy of the responses when the hints are and are not used, respectively; \( \beta_{i2} \) and \( \beta_{i3} \) are item intercepts in the second and the third node, and \( a_{i2} \) and \( a_{i3} \) are the slopes of the item in the second and the third node. Conditional on hint use, the latent variables, the item intercepts, and item slopes in the model for response accuracy can be different. Restricted versions of the model can also be considered in which some of the model elements are the same between the second and the third node.

**Methods**

The set of models described in the previous section were fitted to the English-from-Portuguese data set. First, the initial model was fitted (Equation 5). Second, the model extension with the

![Figure 3](image-url). Tree-structure for responding to an item when hints are available on-demand (\( Y_i \)—hint use, \( X_i \)—response accuracy).
differences in discrimination was fitted (Equation 6). Third, the further model extension with the category-specific parameters \( \delta_n \) was fitted (Equation 9), followed by the final extension with the additional latent variable (Equation 12). In the first four models, the scoring function \( a = [0, 1, 2, 3] \) was used. In addition to that, four multidimensional nominal response models with alternative scoring functions were fitted. Finally, five different IRTree models were fitted with the following specifications on the differences between the second and the third nodes of the IRTree:

1. \( \theta_1 = \theta_2, \ a_{2i} = a_{3i}, \ \beta_{2i} = \beta_{3i} \)
2. \( \theta_1 = \theta_2, \ a_{2i} \neq a_{3i}, \ \beta_{2i} = \beta_{3i} \)
3. \( \theta_1 = \theta_2, \ a_{2i} = a_{3i}, \ \beta_{2i} \neq \beta_{3i} \)
4. \( \theta_1 = \theta_2, \ a_{2i} \neq a_{3i}, \ \beta_{2i} \neq \beta_{3i} \)
5. \( \theta_1 \neq \theta_2, \ a_{2i} \neq a_{3i}, \ \beta_{2i} \neq \beta_{3i} \)

The models were estimated using marginal maximum likelihood assuming a (multivariate) normal distribution of the latent variables in the population. For estimating the first two models (Equations 5 and 6) we wrote an EM-algorithm in R (R Core Team, 2018). For estimating all other models, we used the R-package mirt (Chalmers, 2012). The unidimensional models were estimated using EM-algorithm, and multidimensional models were estimated using Metropolis-Hastings-Robbins-Monro algorithm. For identification purposes, the means and the variances of the latent variables in all models except the first model were fixed to zeros and ones, respectively. In the first model, only the mean of the latent variable was constrained to zero, and the variance was estimated freely.

The models were compared using information criteria that are commonly used in the context of IRT models: Akaike Information Criterion (AIC; Akaike, 1973) and Bayesian Information Criterion (BIC; Schwarz, 1978). Furthermore, to evaluate out-of-sample prediction of the models, two-fold cross-validation was performed. The dataset was randomly split into two equal-sized subsets. Each of these subsets was treated as a testing set, while the corresponding remaining half of the data was used as a training set. All models were fitted to each of the training sets, the log-likelihood was computed for the corresponding testing set. After selecting the best models we fitted them to the full data set to obtain the final estimates of the model parameters, and also to compute the expected-a-posteriori (EAP) estimates of \( \theta \) and \( \eta \) for each person. Furthermore, we fitted two 2PL models and obtained EAP estimates of ability under these models to compare those to the estimates of ability in the selected model. First, we considered a 2PL model for response accuracy which ignores hint use data. Second, we considered a 2PL model for data scored such that the full credit is only given to correct response without hints, and all other outcome are treated as errors.

### Results and Discussion

Table 1 shows the average AIC and BIC across the training data sets and the average log-likelihoods of each of the fitted models across the two testing sets (cross-validation log-likelihood, CVLL) of the English-from-Portuguese data. According to BIC and CVLL the best model is one of the two-dimensional scoring-rule-based models. According to the AIC the best model is the IRTree model in which the abilities, difficulties, and discriminations are different for when the hints are used or not. Given that in the context of IRT models AIC often prefers models that are too complex, we will mainly focus on the model that is chosen by BIC and CVLL. However, we also briefly discuss the IRTree model selected by AIC. In this model, the correlations between the hint-use dimension and the ability dimensions given hints and no hints were .36 [CI: .32, .40] and .05 [CI: .02, .09], respectively. The correlation between the two ability dimensions was .69 [CI: .67,
The correlation between the intercept parameters in the hint-use node \((M = 2.14, SD = 1.46)\) and the intercept parameters in the other two nodes were \(-.10 [Cl: -.29, .10]\) for when hints were used and \(-.56 [Cl: -.68, -.41]\) for when hints were not used. The correlation between the intercept parameters in the second \((M = -2.24, SD = 0.62)\) and the third \((M = -2.17, SD = 0.65)\) node was \(.51 [Cl: .35, .64]\). None of the correlations between the slope parameters were significant \((\alpha_1: M = 2.68, SD = 0.81; \alpha_2: M = 0.23, SD = 0.37; \alpha_3: M = 0.68, SD = 0.23)\).

For the scoring-rule-based models, each of the consecutive model extensions (adding discrimination parameters, adding extra threshold parameters, and adding the second latent variable) resulted in the improvement of model fit in the testing data and in the improvement of BIC, which means that each of these extensions appears needed to better model the joint distribution of hint use and accuracy. Among the different two-dimensional scoring-rule-based models, the model with the initial scoring rule inspired by the signed-residual-time model is not the one having the best values for the information criteria and CVLL. The best-fitting scoring-rule-based model is the two-dimensional model in which in the ability dimension incorrect responses with and without hints are not distinguished from each other in terms of scores.

In the selected scoring-rule-based model, the correlation between the two dimension was equal to \(.13 [Cl: .09, .17]\). That is, there is a significant positive correlation between the ability and hint-use dimensions, but this relationship is very weak (i.e., more able persons are slightly more likely to use hints). The item loadings were higher in the hint-use dimension \((M = 2.15, SD = 0.64)\) than

| Model | npar | AIC   | BIC   | CVLL  |
|-------|------|-------|-------|-------|
| IH\_ < IH\_ \( \times \) CH\_ < CH\_ \( \times \) no \( \alpha \), single \( \delta \), no \( \eta \) | 100  | 275,065 | 275,621 | -137,432 |
| IH\_ < IH\_ < CH\_ < CH\_ \( \times \) single \( \delta \), no \( \eta \) | 198  | 273,548 | 274,649 | -136,867 |
| IH\_ < IH\_ < CH\_ < CH\_ \( \times \) no \( \eta \) | 396  | 241,118 | 243,320 | -120,584 |
| IH\_ < IH\_ < CH\_ < CH\_ \( \times \) single \( \delta \), no \( \eta \) | 96   | 210,563 | 213,322 | -105,273 |
| IH\_ < IH\_ < CH\_ < CH\_ \( \times \) single \( \delta \), no \( \eta \) | 496  | 210,622 | 213,381 | -105,304 |
| IH\_ < IH\_ < CH\_ < CH\_ \( \times \) single \( \delta \), no \( \eta \) | 496  | 210,522 | 213,280 | -105,254 |
| IH\_ < CH\_ < CH\_ < CH\_ \( \times \) single \( \delta \), no \( \eta \) | 496  | 210,653 | 213,412 | -105,327 |
| IH\_ < IH\_ < CH\_ < CH\_ \( \times \) single \( \delta \), no \( \eta \) | 496  | 210,754 | 213,512 | -105,361 |
| IH\_ < IH\_ < CH\_ < CH\_ \( \times \) single \( \delta \), no \( \eta \) | 397  | 211,454 | 213,662 | -105,729 |
| IH\_ < IH\_ < CH\_ < CH\_ \( \times \) single \( \delta \), no \( \eta \) | 496  | 210,975 | 213,733 | -105,543 |
| IH\_ < IH\_ < CH\_ < CH\_ \( \times \) single \( \delta \), no \( \eta \) | 496  | 210,545 | 213,304 | -105,283 |
| IH\_ < IH\_ < CH\_ < CH\_ \( \times \) single \( \delta \), no \( \eta \) | 595  | 210,445 | 213,754 | -105,271 |
| IH\_ < IH\_ < CH\_ < CH\_ \( \times \) single \( \delta \), no \( \eta \) | 597  | 210,416 | 213,736 | -105,267 |

Table 1. Model comparison results for scoring-rule-based models with different specifications (the order of categories for ability measurement is indicated: I—incorrect, C—correct, H.—without hints, H._—with hints; “no \( \alpha \)” means that item-specific discrimination parameters are not included, “single \( \delta \)” means that category-specific thresholds are not included, “no \( \eta \)” means that hint-use dimension is not included) and for IRTree models with different specifications of the second (accuracy when the hint was used) and third (accuracy when the hint was not used) nodes of the tree: npar—number of freely estimated parameters, AIC—average Akaike Information Criterion in the training data, BIC—average Bayesian Information Criterion in the training data, CVLL—average log-likelihood in the testing data. The best model based on AIC is the IRTree model with the different abilities, slopes, and intercepts when hints are and are not used. The best model based on BIC and CVLL is the two-dimensional scoring-rule-based model with highest scores for correct responses without hints, partial scores for correct responses with hints, and zero scores for all incorrect responses.
in the ability dimension ($M = 0.32, SD = 0.12$). This is in line with the findings of Goldin et al. (2013), who found that the individual differences in the tendency to use hints was larger than those in proficiency. Some of the items had very low loadings on the ability dimension (e.g., 0.04 and 0.09 for two of the items), which means that they are hardly contributing to the measurement of ability.

Next, we compare the estimates of ability in the selected scoring-rule-based model with estimates of ability in much simpler models in which hint-use data are either ignored, or all responses with hints are treated in the same way as incorrect responses. The upper panels of Figure 4 show scatterplots comparing the two sets of ability estimates. On the left are the abilities in the 2PL model which ignores hint-use data. They are very close to the abilities in the scoring-rule-based model, which make sense because both models give credit to all correct responses. On the right are the abilities in the 2PL model with the responses with hints coded as incorrect. For a majority of the persons, there is a notable discrepancy between the estimates under the two models. For some persons, however, there is a very strong correlation between the two sets of abilities. These are the persons who never used hints. For those persons who did use some hints, the estimates differ since in one model correct responses with hints get a partial credit while in the other they receive no credit. The lower panels of Figure 4 show how the difference between the estimates of ability in the scoring-rule-based model and each of the 2PL models depends on the estimate of the hint-use latent variable. On the left, one can see that for the persons with low hint-use the estimate of ability is higher for the scoring-rule-based model, while for high hint-use the

![Figure 4. Comparison between the ability ($\theta$) estimates under the selected model (two-dimensional scoring-rule-based method with full credit for correct responses without hints, partial credit for correct responses with hints, and no credit for all incorrect responses) and two-parameter logistic models (2PL) with (1) all correct responses receiving full credit (panels on the left); (2) only correct responses without hints receiving full credit (panels on the right). The upper panels show the relationships between the estimates under the two models. The estimates are a lot more similar when comparing the scoring-rule-based method with the 2PL which ignores hint-use than when comparing it to the 2PL which treats hint-use as incorrect responses. The lower panels show how the difference in the estimates depends on the hint-use dimension ($\eta$): when comparing to the 2PL which ignores hint use for the persons with low $\eta$ the $\theta$-estimate is higher for the scoring-rule-based model and vice verse for persons with high $\eta$; when comparing to the 2PL which treats all hint use as incorrect responses, the opposite pattern is present.](image-url)
estimate of ability is generally higher for the 2PL with all correct responses receiving full credit, but there is also more variance in the difference between the estimates. These differences are due to the fact that correct responses with hints are penalized in the scoring-rule-based method compared to the 2PL. On the right, one can see that the higher hint-use is the larger the difference between the estimates is. That is, persons who use more hints (and also give correct responses) obtain a higher score under the scoring-rule-based methods than under the 2PL with only correct responses without hints receiving full credit.

The differences found between the ability estimates in the selected model compared to the more traditional approach of measuring ability in applications in which hints can be requested pose a question about which way of operationalizing ability is more useful in practical applications. Should only successful independent execution of a task be indicative of ability or is successful completion with some help or guidance also informative? An important question which goes beyond this paper is about the validity of the different ability measures. To answer this question, one would need to, for example, evaluate criterion validity by comparing the correlations between the different ability measures and an external criterion which in the case of our application could be a score on a standardized language test. Furthermore, evaluation of convergent validity by computing the correlations between ability measures obtained from different types of language tasks would help in deciding which measure of ability to prefer.

The exact interpretation of the second latent variable in the model remains an open question. Learners with high values of $\eta$ use more hints than learners with lower values of $\eta$ over and above what can be expected from their ability. However, where these differences come from can be different. For example, some learners might be using hints as a learning tool, while other learners might only give a response when they are very certain of its correctness and otherwise request a hint to avoid any errors. An important topic for further research would be the relationship between the hint-use dimension and other person variables, for example gender, age, and psychological traits. The interpretation of this latent variable also depends on the application at hand, and in particular on the consequences of making an error in a learning system (e.g., whether points are subtracted from a learner, and whether a learner is forced to repeat a lesson instead of moving on) and on consequences of asking for a hint (e.g., whether a learner has to “pay” for the hint in one way or another).

An important limitation of the current study is that the data were collected in a learning context, that is, where ability is changing, but the models assume that ability is constant. The small positive correlation, between ability and the hint dimension can be possibly explained by fact that those students who ask for more hints learn more and thus their average ability over the whole period is higher than students who ask for fewer hints and learn less. It would be interesting to apply the models in settings where no change in ability can be assumed and compare the results in terms of the relationship between the latent variables. Furthermore, the proposed models could be extended with a component that accounts for learning similar to the Additive Factor Model (Cen et al., 2006).

**Conclusion**

In this paper, we have described two different modeling strategies for hint use and response accuracy data from adaptive learning and assessment systems. Starting from rather simple models, we extended them in ways that deemed relevant based on the first empirical data set used for model building and tested these different models in the second empirical data set.

In the application that was considered, the best model was the one which included hint use in the measurement of ability through assigning particular scores to different combinations of hint use and accuracy with all incorrect responses receiving no credit regardless hint use. Furthermore, there were strong individual differences in the learners’ tendency to request hints. However, in a particular application depending on the type of practice items, the type of hints available in the
learning system, and on the consequences of making an incorrect response for the learner within a system, the best way of modeling hint use might be different. Furthermore, making the way in which hint use is incorporated in the measurement of ability known to the learners might also change the way in which they behave and the way in which the data should be modeled.

The modeling strategies explored in this paper have potential for further developments. The scoring-rule-based models can be further explored and modified to account for additional phenomena in the learning data (e.g., strong conditional dependence between item responses) or include a different scoring rule. The IRTree models can also be further explored incorporating different tree structures (e.g., having not only binary nodes, but nodes with multiple outcomes) and multiple item attempts. Furthermore, models can be developed that take into account not only response accuracy and hint use, but also other process data variables in one scoring rule, while at the same time taking into account additional individual differences in learners behavior as indicated by the process data. For example, based on log-file data different solution strategies can be identified and different scores can be assigned to correct responses obtained with different strategies.

Appendix

Derivation of the Scoring-Rule-Based-Model

In order to specify the scoring-rule-based model for response accuracy and hint use, we assume that scores on different items (e.g., items $i$ and $j$) from the same person are independent given ability:

$$S_{pi} \perp S_{pj} \mid \theta_p,$$

which is the traditional conditional independence assumption used in most IRT models. Next, we assume that the score of a person is the sufficient statistic for ability:

$$(X_p, Y_p) \perp \theta_p \mid \sum_i S_{pi},$$

where $X_p$ and $Y_p$ are the vectors of response accuracy and hint use of person $p$, respectively.

Similar to persons, items differ from each other: On some items, people tend to get a high score, while on other items people tend to get a low score. The model can account for such differences between items by assuming that the total score of an item (i.e., $\sum_p S_{pi}$) is the sufficient statistic for an item difficulty parameter $\delta_i$:

$$(X_i, Y_i) \perp \delta_i \mid \sum_p S_{pi},$$

where $X_i$ and $Y_i$ are the vectors of response accuracy and hint use of all persons to item $i$, respectively.

Taken together, these assumptions imply that the joint distribution of response accuracy and hint use on different items can be expressed as follows:

$$f(X_p, Y_p \mid \theta_p) = \prod_i f(x_{pi}, y_{pi} \mid \theta_p),$$

where

$$f(x_{pi}, y_{pi} \mid \theta_p) = \frac{\exp(S_{pi}(\theta_p - \delta_i))}{Z_{pi}},$$
with $Z_{pi}$ representing a normalization factor that makes the probability density function sum to one:

$$Z_{pi} = \sum_{x=0}^{1} \sum_{y=0}^{1} (3x - (2x - 1)y)(\theta_p - \delta_i).$$ (24)

**Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) received no financial support for the research, authorship, and/or publication of this article.

**ORCID iD**

Maria Bolsinova https://orcid.org/0000-0002-8376-5951

**Notes**

1. Note, that for some items the system allows for multiple correct translations, and some typos are recognized as correct.
2. We have switched to the notation with category intercept parameters $\delta_i$, instead of difficulty parameter $\delta_i$, because this parametrization is more convenient computationally, and because it is more common for multidimensional models which are considered in the remainder of this section.
3. Multivariate normality is a common assumption for multidimensional nominal response models. Testing this assumption, evaluating the robustness of the estimation procedure to its violations, and developing estimation tools that do not make such an assumption go beyond the scope of the current paper, but are important topics for further research.
4. Different approaches to model comparison can be considered in addition to or as alternatives to information criteria and cross-validation, for example, root mean squared error, area under the receiver operating characteristic curve, and other measures of accuracy. However, since these methods are not commonly used in IRT, we leave evaluation and application of such methods for further research.

**References**

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, & F. Csaki (Eds.), Second international symposium on information theory (pp. 267–281). Akademiai Kiado.

Aleven, V., & Koedinger, K. R. (2000). Limitations of student control: Do students know when they need help? International conference on intelligent tutoring systems, Montréal, Canada, 19–23 June, 2000. (pp. 292–303).

Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2004). Toward tutoring help seeking. International Conference On Intelligent Tutoring Systems, Maceió, Alagoas, Brazil, 30 August–3 September 2004. (pp. 227–239).

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. Lord, & M. Novick (Eds.), Statistical methods for modeling human dynamics: An interdisciplinary dialogue (pp. 397–479). Addison-Wesley.

Boeck, P. D., & Partchev, I. (2012). Irtrees: Tree-based item response models of the glmnm family. Journal of Statistical Software, Code Snippets, 48(1), 1–28. https://doi.org/10.18637/jss.v048.c01.
Cen, H., Koedinger, K., & Junker, B. (2006). Learning factors analysis—a general method for cognitive model evaluation and improvement. International Conference on Intelligent Tutoring Systems, Jhongli, Taiwan, 26–30 June, 2006. (pp. 164–175).

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06.

Feng, M., & Heffernan, N. (2010). Can we get better assessment from a tutoring system compared to traditional paper testing? can we have our cake (better assessment) and eat it too (student learning during the test)? International conference on intelligent tutoring systems, Pittsburgh PA, 14–18 June, 2010. (pp. 309–311).

Goldin, I., Koedinger, K., & Aleven, V. (2013). Hints: You can’t have just one. *Proceedings of Educational Data Mining 2013*, Memphis, TN, 6–9 July, 2013

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Springer Science & Business Media.

Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods, 48*(3), 1070–1085. https://doi.org/10.3758/s13428-015-0631-y.

Kyllonen, P., & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence, 4*(4), 14. https://doi.org/10.3390/jintelligence4040014.

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/10.1007/s11336-012-9288-y.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. https://doi.org/10.1007/BF02296272.

Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *ETS Research Report Series, 16*(2), 159–176. https://doi.org/10.1177%2F014662169201600206.

R Core Team (2018). *R: A language and environment for statistical computing [computer software manual]*. https://www.R-project.org/.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danmarks Paedagogiske Institut. https://books.google.com/books?id=aB9qlgEACAAJ.

Razzaq, L., & Heffernan, N. T. (2010). Hints: is it better to give or wait to be asked? *International Conference on Intelligent Tutoring Systems*, Kingston, Jamaica, 3–7 June, 2010. (pp. 349–358).

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4 Pt.2), 100.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/aos/1176344136.

Shih, B., Koedinger, K. R., & Scheines, R. (2011). A response time model for bottom-out hints as worked examples. In *Handbook of educational data mining* (pp. 201–212). Routledge.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393–408.

Thissen, D., & Cai, L. (2016). Nominal categories models. In W. J. van der Linden (Ed.), *Handbook of item response theory (I)*, pp. 51–73. CRC Press. https://doi.org/10.1201/9781315374512-7.

Ueno, M., & Miyazawa, Y. (2017). Irt-based adaptive hints to scaffold learning in programming. *IEEE Transactions on Learning Technologies, 11*(4), 415–428. https://doi.org/10.1109/TLT.2017.2741960.

van der Linden, W. J., & Hambleton, R. K. (2013). *Handbook of modern item response theory*. Springer Science & Business Media.

van Rijn, P. W., & Ali, U. S. (2018). A generalized speed–accuracy response model for dichotomous items. *Psychometrika, 83*(1), 109–131. https://doi.org/10.1007/s11336-017-9590-9.

Wang, Y., & Heffernan, N. T. (2011). The “Assistance” model: Leveraging how many hints and attempts a student needs. FLAIRS Conference, Palm Beach, FL, 18–20 May, 2011

Wang, Y., Heffernan, N. T., & Beck, J. E. (2010). Representing student performance with partial credit. *Educational data mining, 335–336.*