# Validating Item Response Theory Models in Simulated Environments

Manuel Hernando, Eduardo Guzmán and Ricardo Conejo

E.T.S. Informática. Universidad de Málaga, Bulevar Louis Pasteur, 35. 29071 Málaga, Spain {mhernando, guzman, conejo}@lcc.uma.es

Abstract. The Item Response Theory is a successful technique generally used in testing systems. Its application in problem solving environments requires the collection of large amount of data. That issue is stressed with ill-defined domains in which the actions that a student could accomplish are difficult to predict. Known IRT models could not be as appropriate as it is desired to that application and we have to explore new alternatives. One of these alternatives is a new family of models called quasipolytomous models of IRT. These models are halfway between dichotomous and polytomous models and require less data than polytomous models being more informative than dichotomous ones. Validating these new models is a very difficult issue in a real environment since student knowledge level is not observable. A simulation environment could help us to verify new models of IRT. Besides, with a simulator we can study different scenarios and observe how our model behaves in them.

**Keywords:** Problem Solving Environments, Student Modeling, Procedural Knowledge Estimate, Item Response Theory.

#### 1 Introduction

Student modeling in problem solving environment is an important issue in the AIED field. Constraint-Based Modeling (CBM) [1] and Cognitive Tutors (CT) [2] are the outstanding approaches in that matter. CBM models are a set of constraints associated to principles in the domain that could be either violated or not, those constraints are related to declarative principles of the domain. CT inferred procedural student knowledge directly from student interactions through a technique called Knowledge Tracing [2] which is based on Bayesian procedures and estimates the probability that a student has learned a certain rule of the domain given his/her actions.

Procedural knowledge could be also inferred by other techniques such as the Item Response Theory (IRT) [3], which is one of the most important strategies of declarative knowledge assessment in testing systems. Our proposal of applying IRT to problem solving environment sets a connection between problem solving environment and

testing, that is, if we make a matching between the elements of problem solving and the elements of testing we can, directly, apply IRT to infer procedural knowledge. In this sense, we model the solution of each problem as a directed graph where nodes are states of the solution path of the problem and edges are transitions between states. Using that representation each node could be understood as a question and each edge as an option in the question.

Our challenge is also to develop an automatic (or semiautomatic) procedure for mining the problem solving path from the logs of students' performance while solving it. This mining process would lead us as well to infer the IRT components which will be used for diagnosing the procedural knowledge. Furthermore, we also want this procedure to be dynamic, that is, the solution path and the inference of IRT components have to be updated dynamically when new logs will be available. Accordingly, the problem graph will include all possible student actions, so when a student completes an action that never was completed by another student it has to be included in the problem graph. These new actions have to be taken into account when the procedural knowledge is assessed, that is, the calibration procedure of the IRT has to be done when new actions are incorporated to the problem graph.

There are, mainly, two families of IRT-based models according to how they update the estimated student knowledge in terms of the student's response: dichotomous and polytomous models. Dichotomous models consider each response as either correct or incorrect whereas polytomous models consider each response individually. These traditional IRT models could not be as good as it is expected dealing with this type of calibrations since dichotomous models are not as informative as we need in this kind of problems and there will be actions with little evidence (maybe actions followed by 1 or 2 students) to polytomous calibration, especially in ill-defined domains where the set of possible actions is very large.

In order to explore new IRT models that fit better with our challenges we have developed a simulation environment in which virtual students (with a known real procedural knowledge of the domain) solve virtual problems (simulating their behavior according to their prior knowledge) and we have compared their estimated knowledge with their real knowledge. In this sense, we have developed a new family of IRT-based models called *quasipolytomous models* which are halfway of dichotomous and polytomous models considering not all possible responses but a subset of them.

## 2 Item Response Theory Models

The IRT is one of the most successful and well-founded strategies for knowledge inference in testing systems [3]. IRT infers and models student performance by means of some probabilistic functions called characteristic curves, the idea is that student's results could be explained by a set of non-observable factors (for instance, the knowledge level).

There are a lot of IRT models, based on how the models update the estimated student knowledge in terms of his/her response they could be [4]:

- Dichotomous models: Each response is considered as either correct or incorrect.
  When a student selects an option in a question test, his/her estimated knowledge is
  updated according to whether the option selected is the correct one or if that is other. These models require only a characteristic curve per item that represents the
  probability that a student with a certain knowledge level answers it correctly. This
  characteristic curve is called item characteristic curve (ICC).
- Polytomous models: Each possible response has its own characteristic curve called
  operating characteristic curve (OCC) [5], which expresses the probability that a
  student select that answer [6]. The student estimated knowledge is updated by
  means of the OCC related to the selected response.

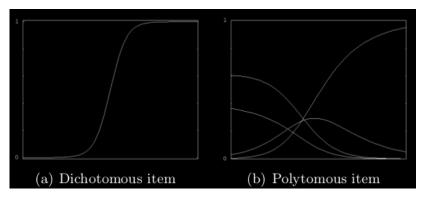


Fig. 1. Characteristic curves of an item under dichotomous and polytomous models

Polytomous models are more informative than dichotomous ones since they take into account each possible response independently instead of considering each answer as correct or incorrect.

Figure 1 shows the curves of an item of both the dichotomous and polytomous models of IRT. The dichotomous model is shown in Figure 1(a) and has only the ICC which is the probability of answering correctly this item (y-axis), given a certain knowledge level (x-axis). The polytomous model is presented in Figure 1(b). Each item choice has its own characteristic curve which is the probability of choosing this choice (y-axis), giving a certain knowledge level (x-axis).

The most popular proposals for modeling the dichotomous characteristic curves are the logistic models, which use logistic functions. These models could be classified, considering the number of parameters that the function has. According to this classification there are 3 kinds of logistic models: 1PL, 2PL, and 3PL, with one, two and three parameters, respectively. A generic 3PL ICC of an item  $X_i$  is defined as follows:

$$P(X_i|\theta) = c_i + (1 - c_i) \frac{1}{1 + e^{-Da_i(\theta - b_i)}}$$
 (1)

Where D is a parameter introduced to fit the curve similar to the normal curve, parameter  $a_i$  is the discrimination, parameter  $b_i$  is the difficulty, parameter  $c_i$  is the guessing parameters of the item  $X_i$ , and  $\theta$  is the knowledge level.

There are a lot of polytomous models of IRT. In our work we have considered the IRT-model proposed by Thissen and Steinberg for multiple-choice items [7]. In this model besides the observable categories (selectable choices) there is another non-observable and latent category called "don't know" (DK) that expresses the probability that a student does not know how to answer the item. Each observable category has a portion of the category DK included since students who do not know will select an observable category. The formula of each observable category is exposed below,  $X_i$  represents the response to the item i and h is the category selected:

$$P(X_i = h|\theta) = \frac{e^{a_h \theta + c_h}}{\sum_{k=0}^{m_i} e^{a_k \theta + c_k}} + d_h \frac{e^{a_0 \theta + c_0}}{\sum_{k=0}^{m_i} e^{a_k \theta + c_k}}$$
(2)

category 0 is the non-observable category DK and  $d_h$  is the portion of that category included in each observable category h. The parameters denoted by a reflects the order, as well as discrimination, for the categories, and the parameters denoted by c reflect the relative frequency of the selection of each alternative.

## 3. Introducing IRT in problem solving

The application of IRT to problem solving environments requires a polytomous model since each student action should be taken into account; a dichotomous model only would be able to establish if an action is correct or incorrect. However, some actions could have little evidence and IRT calibration could be not as accurate at it is expected. For that reason, we have developed a new family of models of IRT called quasipolytomous models of IRT which are on the halfway between dichotomous and polytomous ones.

Quasipolytomous models consider not all choices as independent but only those that have enough evidence. For instance, let us consider an item witch 20 choices (what is usual in problem solving environment if we include all student actions), if 8 of them have been selected only by 1 or 2 students they do not offer us enough evidence to do a polytomous calibration. Instead of doing it, we consider these 8 choices as a simple choice reducing the number of OCCs from 20 (one per choice) to 13 (12 individual choices and an extra choice that group the other 8) that have, all of them, evidence enough to do an IRT calibration.

In testing systems there could be items with a lot of choices too, let us consider figure 2 in which the number of choices is 120 since we have to take into account the permutation between these 5 elements.

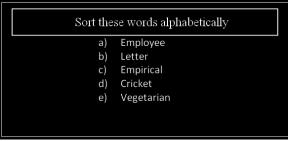


Fig. 2. An item with 120 choices

### 3 Simulation environment

In order to verify quasipolytomous models we have developed a simulator in which virtual students have a real (prior) knowledge assigned and they have to solve some virtual problems according to their knowledge level. Once the students solve the proposed problems, their knowledge is estimated by means of a quasipolytomous model of IRT and then, these estimations are compared with the students' real knowledge.

## 3.1 Virtual problems

In this simulation environment, a virtual problem is represented as a collection of items; each item is a state of problem solving path with a certain number of possible transitions to other states. These transitions are the choices of the item.

While students are solving problems, new students' actions could appear in the system and they need to be included in the model. For that reason, virtual problems are not static entities but they can change during students' interactions.

At the beginning, each problem has only the ideal solution path, that is, those nodes that are part of the ideal solution. Those nodes are modeled by dichotomous items according to the equation 1.A characteristic curve, the ICC, is enough to model this kind of items. Students who do not answer correctly the item could, with some probability, make a new action at this step of problem solving. Then the opposite curve (i.e. 1-ICC) is branched into two curves changing the original dichotomous item to a polytomous one. According to the former explanation, students' actions could provoke the addition of new nodes to the problem graph.



Fig. 3. Addition of new curves to an item

Figure 3 shows how new curves are added to an item. Firstly we have the opposite ICC curve which expresses the probability that a student, given his/her knowledge level, will answer the item incorrectly. This is curve marked as a in the figure. Curve a is branched into curves b and c when the action represented by curve c is added to the model; finally, curve c is branched again and then curves d and e are added to the item. The item was, at the beginning, a dichotomous item with two possible responses; then a new action was included in it and, as a consequence, the wrong curve was converted in other two curves. Again a new action was added to the model and the item changes to a 4-choices item.

#### 3.2 Virtual students

A virtual student is an entity that has a real (prior) knowledge associated and is able to solve problems according to it. The idea is that, depending on its knowledge level, the student could go on through the problem graph by completing an action or other. Further, they could accomplish new actions with a certain probability.

A student will select an action of the state according to his/her knowledge level and the characteristic curve since these curves are probabilistic functions of the knowledge level. Figure 4 shows an example of a virtual student selecting a choice in a 4-choice item. That student will select each choice with a probability of 0.2, 0.38, 0.18 and. 0.24 respectively. When a non-correct action is selected, there will be a probability that the student will accomplish a new action and, therefore, this curve will be branched into other two.

Since real knowledge of virtual students is assigned before simulation, we can choose what population distribution we want in each experiment. It is an important feature of the simulation since we can study the impact of different kind of population in order to validate new models in all cases or only in some of them.

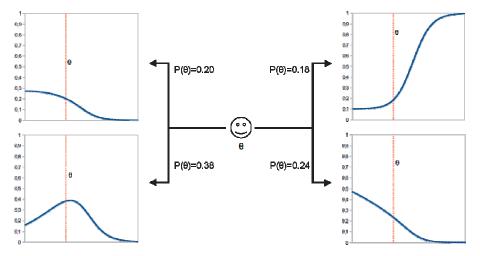


Fig. 4. Virtual student selecting a choice

## 4 Experimentation

In order to verify our model in a simulated environment, we have conducted some experiments with virtual students and virtual problems. The experiments have been accomplished to verify the model with different population distributions as well as different models of IRT and different number of problems.

# 4.1 Experiment description

The experiments accomplished with our simulator were able to compare scores offered by quasipolytomous models of IRT in different situations. Each experiment has been accomplished 30 times in order to reduce the impact of anomalous data. The number of virtual students in each experiment was 1000.

In order to get confidence results we have not done the calibration IRT phase, instead of that, we have estimated knowledge level with the known curves of each model. Polytomous models could not be as accurate as expected including the calibration stage since some item choices could have not enough evidence to be calibrated properly. The student estimated knowledge level is calculated using the formula of the equation 3, the probability of having a specific knowledge level given the steps followed by the student solving the problem is calculated multiplying the probability of selecting each step given the knowledge level.

$$P(\theta|s_1, s_2, \dots, s_n) = \prod_{i=1}^n P(s_i|\theta)$$
(3)

In our experiments, virtual students have solved two problems with 10 items. We have accomplished mainly three experiments varying the population distribution, the probability of generating new actions and the item difficulty respectively. In our experiments, we have compared accuracy of knowledge level estimation obtained by a quasipolytomous model with those obtained by polytomous and dichotomous models with the same data.

The accuracy of knowledge level estimation was calculated using the next formula, where  $\theta$  is the real knowledge,  $\theta^*$  is the estimated knowledge, and N the number of students:

$$d(\theta, \theta^*) = \frac{\sum_{i=1}^{N} (\theta - \theta^*)^2}{N}$$
(4)

We have chosen four types of population distribution to accomplish the experiment. First, a normal distribution of population was selected; the probability of having a knowledge level was centering in the middle of knowledge range, since that range was [-3, 3]; the distribution was centered in 0. The second population was a uniform one, in which the probability of having a knowledge level is equally distributed. The other two populations were a low-level and a high-level population, which were normal distributions centered in small and high values respectively. Figure 5 shows the

knowledge level distributions used in our experiments. From left to right they are the normal, the uniform, the low-level, and the high-level distribution respectively.

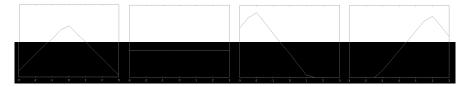


Fig. 5.Students' knowledge level distributions used in the experiments

The impact of the generation of new actions was also studied in our experiments. To this end, we have conducted experiments varying the percentage of adding a new action giving it values of 1%, 1.5%, 2%, and 2.5%. More than 2.5% of adding new actions could lead to a very large number of curves in the model.

Finally, we have also considered the difficulty of the items. We have conducted an experiment changing the value of this parameter. The difficulty value is calculated according to a normal distribution centered in a certain value, which is the difficulty average. We have done our experiments with values of -1, 0, and 1, respectively.

#### 4.2 Experiment results

Experiments conducted suggest that polytomous models of IRT perform more accurately than dichotomous ones. This result is not surprising, since polytomous models are more informative. Experiments also show that quasipolytomous models of IRT are not as accurate as polytomous ones but more accurate than dichotomous ones. Besides, results obtained by quasipolytomous models are very similar to those obtained by polytomous ones.

Table 1 shows results obtained with different population distribution. In all cases accuracy obtained by polytomous models are better than obtained by dichotomous and quasipolytomous models. These differences are higher in the normal distribution and lower with a low-level population.

Distribution	dichotomous	quasipolytomous	polytomous
Normal	4.153118	0.7957639	0.6714305
Uniform	3.841433	0.6284451	0.5397757
Low-level	2.290098	0.4186174	0.2723583
High-level	3.784738	0.8671618	0.8108576

Table 1. Accuracy of IRT models changing the population distribution

We can see in the former table that differences between results obtained by quasipolytomous models and polytomous ones are not significant.

Our second experiment conducted studied how affects the probability of adding a new action. To this end, we changed the percentage of adding a new action from 1%

to 2.5%. Table 2 shows those results, when we increase the probability of adding a new action the accuracy of dichotomous and quasipolytomous models gets worse since but the accuracy of the polytomous models gets better since they obtain more precise information.

Table 2. Accuracy of IRT models changing the percentage of adding a new action

% new action	dichotomous	quasipolytomous	polytomous
1.0%	3.291380	0.6609889	0.6126646
1.5%	3.489956	0.6695576	0.5845097
2.0%	3.613464	0.6778118	0.5593333
2.5%	3.674588	0.7016298	0.5379146

Finally, we have compared results obtained by these models varying item (solving path step) difficulty average. The student knowledge level used in our experiment is in the range [-3, 3]. We have considered average item difficulty of -1, 0, and 1. An item with difficulty of  $b_i$  will be answered correctly by a student with knowledge level of  $b_i$  with a probability of 0.5. Table 3 shows the results of this experiment, the dichotomous models have a better behavior when the difficulty is below 0, however quasipolytomous and polytomous models get better results when the difficulty increases. Polytomous and quasipolytomous models accurate better when the item difficulty is higher since it allows the model to have more curves to estimate the student knowledge level.

**Table 3.** Accuracy of IRT models changing the item difficulty average

Difficulty	dichotomous	quasipolytomous	polytomous
$b_i$ = -1	2.364223	1.2011291	1.0558448
$b_i = 0$	3.131082	0.5410807	0.4450589
$b_i = 1$	5.056735	0.2902812	0.2199130

#### 5 Conclusion

In this paper, we have validated a new approach that uses the Item Response Theory, a well-founded theory generally used for declarative knowledge estimation in testing systems, to infer procedural skills in problem solving environments. To do that, we have developed a new model of IRT, the quasipolytomous model. This model is halfway between dichotomous and polytomous models being more informative than dichotomous models and needing less amount of data than polytomous ones.

This verification could be difficulty accomplished in a real environment, since we need to know the prior knowledge level of students to measure the estimation accuracy. This knowledge level, however, is a latent trait that is not observable. In addition,

we needed a controlled environment where the students' performance was not biased by external factors. For all these reasons we have developed a simulation environment.

Using a simulation environment we can choose the nature of the population in order to study how well the model performs in different situations and with different students' samples. We also can decide the difficulty of the steps of the problem (i.e. items): before any calibration, the simulator is able to decide what items are more difficult and what are easier.

Other advantage of using a simulation environment is that we can repeat each experiment in order to reduce the impact of anomalous data. Furthermore, we can present the problems to a large number of students, which could be difficult in a real environment.

Regarding the quasipolytomous model of IRT, our experiments show that its application is useful in problem solving environment (besides any kind of procedural task, and in declarative domains for inferring declarative knowledge), especially if we work with ill-defined domains in which the amount of possible new actions is very large. Quasipolytomous models of IRT offer similar estimations as polytomous models but need less data. Besides, quasipolytomous models of IRT are more informative than dichotomous ones since they collect data from correct and incorrect responses.

**Acknowledgements**: This work is part of DEDALO project which is financed by the Andalusian Regional Ministry of Science, Innovation and Enterprise (P09-TIC-5105).

### 6 References

- Mitrovic, A., Martin, B., Suraweera, P.: Intelligent tutors for all: The constraint-based approach. IEEE Intelligent Systems 22 (2007) 38-45
- 2. Anderson, J.R., Corbett, A.T., Koedinger, K.R., Pelletier, R.: Cognitive Tutors: Lessons learned. The journal of the Learning Sciences 4 (1995) 167-207
- 3. Embretson, S.E., Reise, S.P.: Item response theory for psychologists. Lawrence Erlbaum, Mahwah (2000)
- Guzmán, E., Conejo, R., de-la Cruz, J.L.P.: Adaptive testing for hierarchical student models. User Model. User-Adapt. Interact. 17 (2007) 119-157
- 5. Dodd, B.G., De Ayala, R.J., Koch, W.R.: Computerized adaptive testing with polytomous items. Applied Psychological Measurement 19 (1995) 5-22
- Guzmán, E., Conejo, R.: A model for student knowledge diagnosis through adaptive testing. In: In Proceedings of 7th International Conference Intelligent Tutoring Systems, ITS2004 Brazil, Springer-Verlag (2004) 12-21
- Thissen, D., Steinberg, L.: A response model for multiple choice items. Psychometrika 49 (1984) 501-519