AACE
Association for the Advancement of Computing in Education
Celebrating 25 Years of Service to the Educational Technology Community

PO Box 1545
Chesapeake, VA 23327-1545 USA
Phone: 757-366-5606
Fax: 703-997-8760
E-mail: info@aace.org
www.aace.org

CERTIFICATE OF PARTICIPATION

Please allow the following to certify that Martin Lesage of the Université du Québec à Montréal presented: “Interfaces Modelization in Computerized Adaptive Testing (CAT) for Learning Optimization,” and “Extending Moodle Functionalities to Adaptive Testing Framework” while attending the E-Learn 2007 World Conference held October 15-19, 2007 in Quebec City, Canada.

E-Learn is the 12th Annual World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, sponsored by the Association for the Advancement of Computing in Education (AACE). AACE is a nonprofit, international organization whose purpose is to advance the knowledge and quality of learning and teaching at all levels through the encouragement of scholarly inquiry related to information technology and education and the exchange of research results, developments, and their applications through publications and conferences for its members.

This annual conference serves as multidisciplinary forum for the discussion and exchange of information on the research, development, and practitioners to learn about the best practices/technology in education, government, healthcare, and business.

Participation in E-Learn 2007 included the Keynote, Invited, Paper, SIG Discussions, Interactive Sessions, Poster, and Corporate Demo Sessions of this international forum. All presentation proposals are reviewed and selected by a respected international Program Committee, based on merit and the perceived value for attendees.

If any further information is required, please contact the AACE office.

Sincerely,

Tracy Jacobs

Tracy Jacobs
E-LEARN /AACE Representative

Extending Moodle functionalities to adaptive testing framework

Komi Sodoké, Roger Nkambou, Gilles Raîche, Martin Riopel, Martin Lesage
Université du Québec à Montréal (UQAM) C.P. 8888, Montréal (QC)
Canada
sodoke.komi@courrier.uqam.ca

Abstract: E-learning has advanced considerably in the last decades allowing the interoperability of different systems and different kinds of adaptation to the student profile or learning objectives. But, some of its aspects, such as E-testing are still in their early age. As a consequence, most of the actual E-learning platforms only offer basic E-testing functionalities. In addition, in most those platforms, the tests are in the traditional format despite their known limitations and precision problems. However, by making efficient use of well known techniques in artificial intelligence, existing theories in psychometry and standards in E-learning, it could be possible to integrate adaptive and more informative E-testing functionalities in the actual E-learning platform. Some experiments have been done with the Moodle platform. In this paper, we will present the some of principles, the architectural elements and the algorithms used.

1. Introduction

Although assessments and testing are important parts of the learning life cycle, in most E-learning platforms, E-testing functionalities remain in the early development stage (Hage & Aïmeur, 2005). One of our main research objectives is to find efficient solutions to those E-testing specifics challenges and their integration into actual E-learning platforms. In this paper we will focus on the integration of adaptive testing functionalities in the Moodle platform. Moodle is chosen because of it's and open source and popular E-learning platforms. This paper will be organized as follow. First we will present the limitations of the classical way of assessment and the proposed solutions witch is adaptive testing. Second, we will present the models and tools used to achieve the target goal. The developed system and some of the element of its architecture will be presented followed by the results and discussion.

2. From classical to adaptive testing

In their traditional format, tests or assessments are the same for all the examinees. They usually have a predetermined duration and fixed number of questions that have various levels of difficulty. The examinee's mark, which is generally the weighted sum of the scores obtained in all the questions, is used as competence criterion or to infer the examinees ability level in the concerned field. This testing format has known limitations and poses various problems of reliability (Wainer, 2000). For instance, a high skill examinee could face some very easy items during the test or on the other hand, a low skill examinee could face some very hard items. In both cases, that situation could lead the concerned learner to a lack of challenge or motivation that could have a substantial impact on his test outcome. Second, the precision of the test is not the same in the whole range of abilities scales especially in the extreme points of the scale. Third, it's easy for the examinees to cheat because they all have the same test with the same questions in the same order. Finally, the mark doesn't give enough information about the student answer vector. But this could help to determine aberrant responses patterns or a gaming behavior from the leaner; features that are important especially for diagnostic testing. Despite these problems, in most of the E-learning platforms, the tests often have the traditional testing format and they are a simple online "paper and pencil" version.

Question and test adaptation has been proposed as solutions for the traditional test problems because the tried to adapt the test to the learner ability and profile to better estimate his proficiency. So, some attempts have been made in order to integrate adaptive items or assessment functionalities in existing E-learning platform. For instance, Moodle offered low adaptive items functionality based on the IMS QTI specifications. So, its adaptive items can change their appearance, scoring or both in response to each of the learner attempts. Also, relevant work has been done on exam question recommender systems (Hage & Aïmeur, 2005). But, in both cases, the sequence of the items in the test didn't change according to the learner ability or profile. The Online Adaptive Testing (**OAT**) we intended to implement is a personalized test that is adapted to the examinee's ability level, his profile and other preferences in order to better estimate his proficiency. So, that adaptive testing, the first question, the subsequent questions in the test or the overall sequence of the questions as well as the end of the test can vary from one examinee to another. In addition, the overall mark is not the focus in adaptive testing, it is rather the specific answer given by the examinee to each question that is considered, as well as the information that could be deduced or inferred based on the success or failure of the question.

The design and implementation of an OAT platform requires taking into account some challenges. Some of them consist of finding computational models allowing to estimate and to compare in an objective way the proficiency of learners that received different questions during a test. These models must also offer mechanisms to initiate the learner model, to determine the first item to be presented and to manage the whole sequence of questions during the test. The models must have mechanisms to manage the dialog between the system and the learner in order to get the most informative test according to the learner proficiency and profile. The platforms have to be build based on actual standard in E-learning to ensure its integration and interoperability with existing E-learning platform.

By taking into account these requirements and specifications, previous attempts to implement computerized adaptive testing and artificial intelligence formalisms, the tools and models retained are: the item response theory, the Bayesian network and the IMS-QTI standard.

3. Tools and models used

3.1 Item response theory (IRT)

Item response theory is a set of related psychometric models that provides a foundation for scaling persons and items based on responses to assessment items (Wainer, H. 2000). The person parameter usually is the proficiency or cognitive ability within a specific domain and it is represented by the Greek letter θ. A question in a test is a simple example of an item.

Much of the literature on IRT focuses on its models. Those models are usually functions relating person parameter θ and item parameters to the probability of a discrete outcome, such as a correct response to that item. Among the available models, we use the three parameters logistic (3PL) model because its give us enough tools for our implementation. Those three item parameters are the discrimination (**a**), the difficulty (**b**) and the pseudo-guessing (**c**). The later represents the chance for a low level examinee to find by guessing the correct response to the item. The conditional probability for a person with an ability θ to get a correct response to an item i (a_i, b_i and c_i) is:

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

where D is a constant which value is 1.701.
IRT provides strategies for:

- Estimating the learner ability θ (Baker, F. 2001) and the standard error (S_θ) of the estimated ability.
- Estimating the item parameters from data (Baker, F. 2004).
- Ascertaining how well the data fits a model, for instance the L_z misfit indices.
- Investigating the psychometric properties of assessments.

For the implementation, an adaptive version of the MAP method is used to estimate the learners' abilities. It ensures to have an average good value of θ within a good response time.

The main reason why IRT was chosen as a model is that the estimated value of the learner proficiency is independent from the items used for the computations. So the estimated value can be used to compare in an objective way the proficiency of learners that received different questions during a test. IRT also provide the concept of item and test information. That information is estimated for an item or a test at a given value of θ. The item information can be used to select the optimal item to administer to the learner.

In addition, IRT has been successfully used in some computer adaptive testing system [6]. IRT is a data oriented model and different studies (Woolf and al., 2005) show its advantages in term of implementation complexity, predictive power and independence from a subjective expert appreciation. However when a learner ability is known in a concepts, IRT doesn't offer a mechanism to infer his proficiency in related concepts. According to cognitive science theories, learners need not only master specifics concepts, but also the relationship, such as similarity, difference, aggregation, etc. between these concepts. Since it's hard to evaluate the learner for all of those previous aspects, the Bayesian network comes as a handy complement of IRT for that purpose.

3.2 The Bayesian networks (BN)

Formally, Bayesian networks are directed acyclic graphs whose nodes represent variables, and whose arcs encode the conditional dependencies between the variables. The nodes can represent any kind of variable, be it a measured parameter, a latent variable or a hypothesis. Efficient algorithms exist that perform inference and learning in Bayesian networks. Then, the BN are a complete model for the variables and their relationships, it can be used to answer probabilistic queries about them. For example, the network can be used to find out updated knowledge of the state of a subset of variables when other variables (the evidence variables) are observed.

By using the IRT it's possible to get the value of the probability for a learner to know specifics concepts. Those values can be added to BN nodes representing the concepts of the domain and used for different inferences. The BN are also to manage uncertainty in the student modeling (Conati and al., 2002).

3.3 IMS-QTI

The IMS Question & Test Interoperability (QTI) specification describes a data model for the representation of question (assessmentItem) and test (assessmentTest) data and their corresponding results reports. Therefore, the specification enables the exchange of this item, test and results data between authoring tools, item banks, test constructional tools, learning systems and assessment delivery systems[1].

For the implementation, the questions are physically stored as an IMS compliant item. Each item also has a sharing rights such as view, use, modify among the authors of the questions, the group they belong to and the other users.

4. The platform mains models

The framework developed is used to extend Moodle in order to make them administer adaptive testing. Then, Moodle can be more adaptive by building a model of the goals, preferences and knowledge state of each learner and use this model throughout the interaction with the learner in order to adapt to the test his ability level. That student model contains for each learner his ability or misconception in different domains and will be used for diagnostic purposes. The model can be consulted by the learner's teachers and download it as XML file for simulations. Consequently, the platform has two mains models: the domain knowledge model and the student model.

The domain knowledge model is elaborated from a learning perspective of the concepts in the concerned domain. For that purpose, taxonomy of the concepts that will be evaluated in that domain is defined. That taxonomy is used to create categories which are physical representations of the concepts. Each category is used to store the items relevant to evaluate the concerned concepts. In addition, an IMS compliant manifest file is placed in the root folder for each domain to describe it content. The figure 1 shows a simple domain knowledge representing the sample the concepts taught in a data base course. The granularity level depends on the domain and the goals.

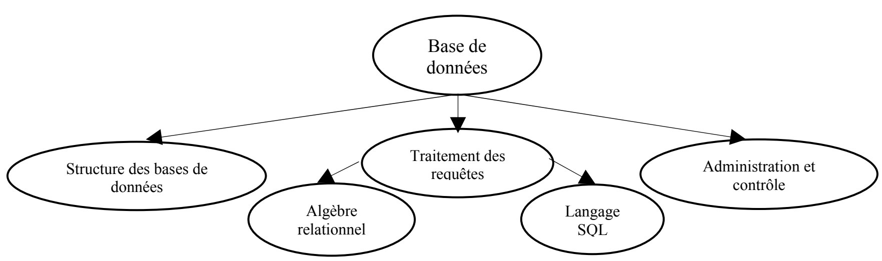

Figure 1. Concepts taught in a data base course

[1] http://www.imsglobal.org/question/

The student model mainly contains the learner's cognitive state. It's made up with the long-term knowledge state, the short-term knowledge state and the episodic memory.

The short-term knowledge state is a temporary structure that principally holds and describes the learner estimate proficiency θ, its standard error (S_θ) and the misfit of the answer pattern (L_z) for the current evaluation session. Their values are updated at each response given by the learner and they are used to select the next item during the test. At the end of the evaluation session the contents of the short-term knowledge state is used to update the long-term knowledge state. That long-term knowledge state represents the learner knowledge state as interpreted through the results of all of the evaluations that he had taken. It is represented as an overlay (Carr and Goldstein, 1977) of the domain knowledge. A Bayesian network is used for its implementation. Each node of the network represents a specific concept of the concerned domain and it expresses the acquisition probability by the learner. The episodic memory stores the traces of the learner evaluations and his actions during the evaluation sessions. It is physically stored as a log file.

5. Element of the platform architecture and their interaction

For the implementation of the framework and the integration into Moodle, two systems were developed.

The first system is a client utility software named PersonFitClient which is used by the tutors to create assessment items. Each item also has a sharing rights such as view, use, modify among the tutors, the group they belong to and the other users; as well as it specific domain taxonomy concepts. The purpose is to build big item pools with different contributors. Among other available functionalities, we have the management of the local items bank, the ability to load and download items into or from the item bank located on the server. The creation of adaptive testing and its simulation using instance of learner's models that can be download as XML files from Moodle.

The second system is based on substantial changes that have been made to Moodle quiz module. These changes enabled us to manage the shared or distributed IMS compliant items pool created by the client software and to display the items in adaptive format. Finally, we integrated some of the functionalities of the client software into Moodle in order to administer OAT to learners.

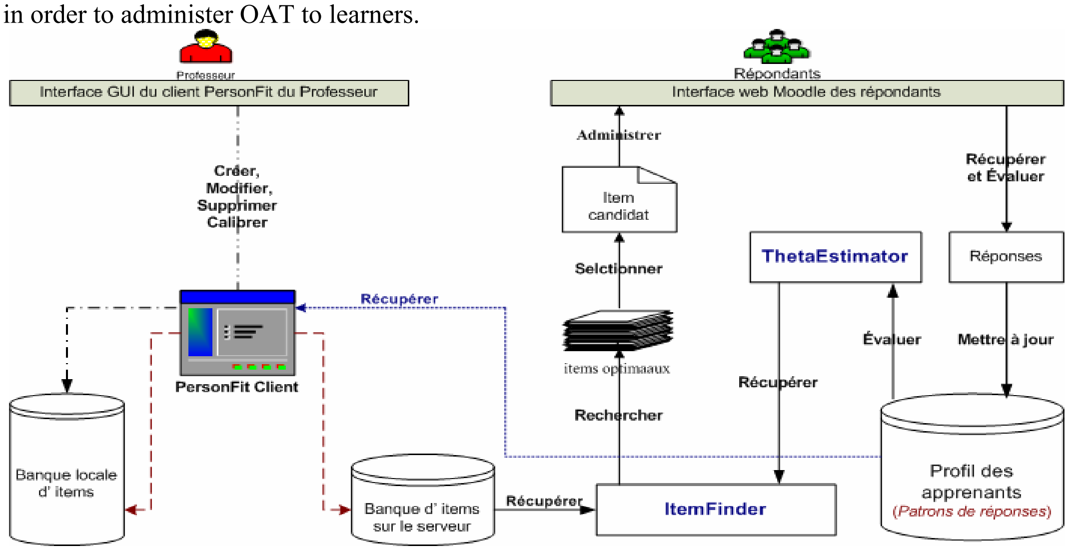

Figure 2. Architecture d'interaction entre des composants de PersonFit et Moodle

The ThetaEstimator is mainly used to estimate the value of the he learner estimate proficiency θ, its standard error (S_θ) and the misfit of the answer pattern (L_z).

The ItemPresentator is used to present the item to the learner. The Figure 4 is a screen shot of the ItemPresentator showing a multi-choice question. At the top of the windows we have the body of the question witch in that case is a simple image. At the bottom, the answer choices are presented. All of the navigation buttons are disabled because this item is presented inside an adaptive test. So, it is the tutor in the right frame of the window who takes the control of the test questions sequence. The, the learner can only give, choose or select his answer or ask the tutor for help.

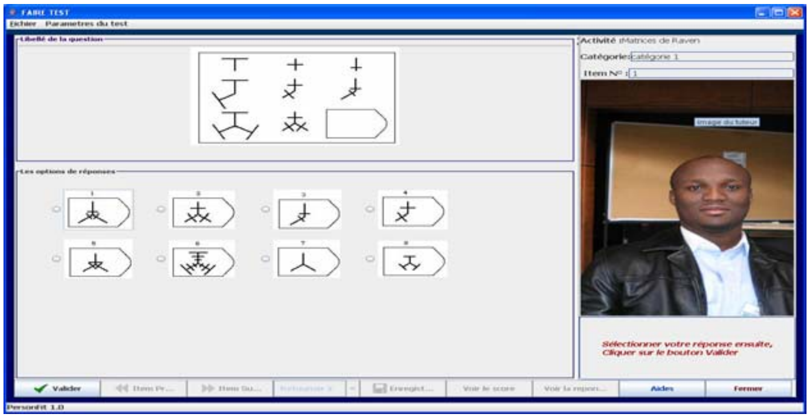

Figure 3. Question presentation interface

The ItemFinder is mainly used to select optimals items to administer to the learner for a given value of their ability. It the ItemFinder get the different values estimated by the ThetaEstimator and select in the item bank on the server a group of optimal items according to those values. After that, the item that will be administered to the learner is randomly selected in that group. That helps to prevent that the learners with the same ability to always get the same item during the test.

6. Adaptive testing algorithm used

Figure 1 presents a simplified flow chart illustrating the adaptive testing algorithm. Most of the steps in that algorithm involves the participation of the ThetaEstimator , ItemFinder , ItemPresentator.

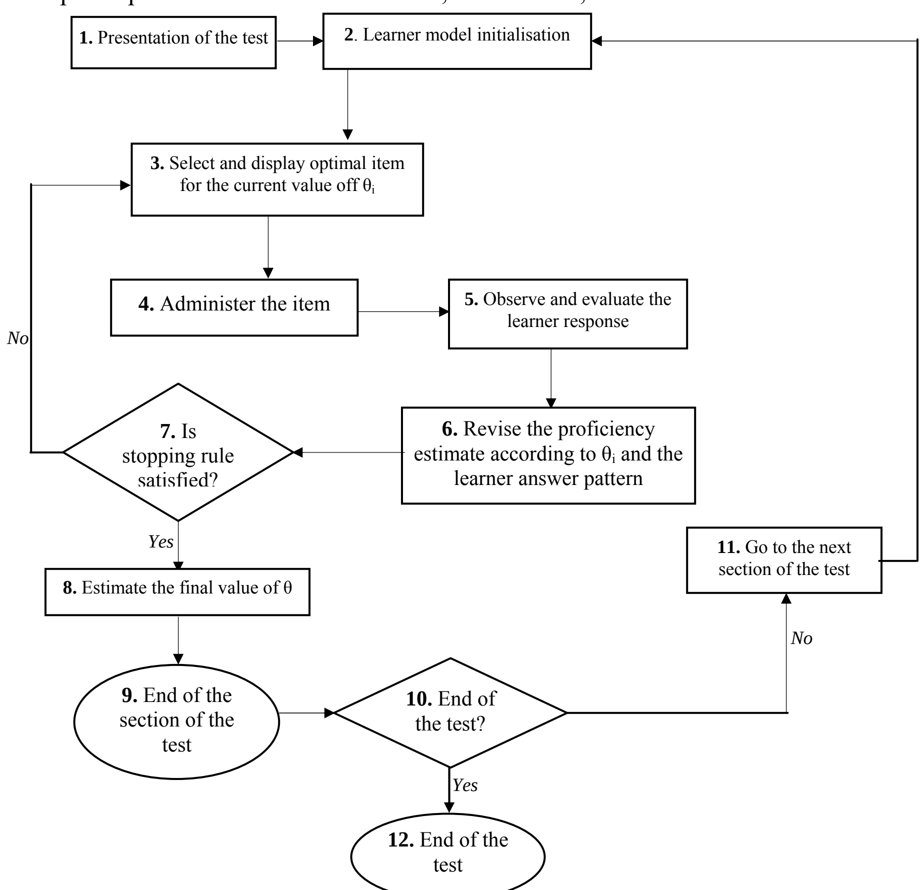

Figure 1. Adaptive testing algorithm flow chart

The step 2 of the algorithm is about the learner model initialization. It aims to estimate the initial value of the learner proficiency named θ_i at the beginning of the current evaluation session. That initial or prior estimated value θ_i is used to select the first item of the adaptive test. If the learner proficiency in the field concerned by the evaluation or some of the concepts related to that field are known, θ_i is obtained by using an inference in the learner long-term knowledge state. But, if that's not the case, an average proficiency value based on the estimated proficiency of other learners in his group in that domain is used. If the information about other learners in his groups is not available, the value of θ_i is set to 0.

At step 3, the strategy used to select the next item consists of choosing the item that will give the maximum information for the learner's current estimated proficiency θ. It would be inefficient to search the entire item pool to compute the information given by each item for a current value of θ and select the most informative one. In practice, an information table is used. This table contains the list of the items ordered by the amount of information provided at various levels of θ [6]. Other constraints relating to a better coverage of the pedagogical contents of the domain and a minimization of some item's exposure are also taken into account. Finally, the episodic memory is checked to prevent the administration of the same questions during different evaluations.

At step 6, the learner proficiency estimate is revised using the adaptive version of MAP.

At step 7, a fixed value of the standard error is used as the stopping criterion. In case of non convergence of θ, the test is stopped after a fixed number of items.

6. Results and discussions

Experiments based on real data have been made in order to determine the accuracy of the adaptive testing implemented in PersonFit. We use the data from the "English as second language classification" test administered in Québec's colleges named TCALS. The non adaptive version of TCALS contains 85 questions and is administrated to thousands of examinees. It gives data file containing thousands of answer vectors coming from a real assessment situation. An experiment of an adaptive version of the TCALS using a sample of 515 examinees taken from the data file has been made. First, we would like to check if the initial value of θ could have an impact on its final estimated value. Second, we want to see the number of items needed to reach the convergence of θ. Then, for the experiments, an initial value of the θ_i is set to -3. At each administered item, the average value of θ_i for all the 515 examinees is graphed giving us the dotted curve beginning at -3. The same procedure is repeated with an initial θ_i value of -2 and so on up to 3. The figure 2 shows the results of the experiment. It represents the estimated value of the ability θ for different number of items.

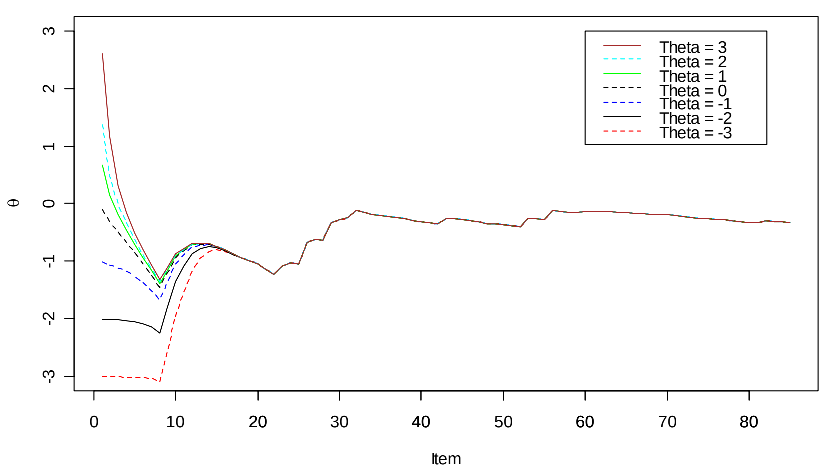

Figure 2. Real data experimentation result

Figure 2 shows that beyond the 15th administrated item, all the curves merged. Then independence from the value of θ_i is obtained. In addition, after the 55th item, there is no significant variation of the value of θ and administration of additional items does not bring us more information. The convergence is then obtained. Then, the adaptive version of TCALS only needs 55 items to estimate the learner ability.

7. Conclusion

Efficient use of IRT combined with Bayesian networks allows the implementation of a functional platform to administer an adaptive testing. That platform is integrated to Moodle by extending it functionalities to manage an independent IMS compliant items bank. Different engines were also added to gather and evaluate the learner response, estimate his ability, select and administer the most informative question. The results obtained by using the proposed adaptive testing algorithm on real data show how this algorithm is efficient. The outcomes of these tests will be used for a diagnostic purpose by using the learner model that contains his ability or misconception in different domains. The diagnostic result can be presented to the learner teachers or used by the platform to recommend a specific leaning path through the online learning material in Moodle to the concerned learner.

References

Baker, F. (2001). *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland

Baker, F. (2004). *Item response theory – Parameter estimation techniques*. New York: Marcel Dekker.
Carr, B. & Goldstein, I. (1977). *Overlays: A theory of modelling for computer aided instruction*, Technical Report AI Memo 406, MIT, Cambridge, MA.

Conati, C. & Gertner, A. & Vanlehn, K. (2002). *Using Bayesian networks to manage uncertainty in student modeling*, Journal of User Modeling and User-Adapted Interaction, volume 12 (4), pages 371-417, 2002.

Hage, H. & Aïmeur, E. *Exam Question Recommender System*. Proceedings AIED (2005)

Wainer, H. (2000). *Computerized adaptive testing: A primer*. Mahwah, NJ: Lawrence Erlbaum Associates.

Jonsson, A., Johns, J., Mehranian, H., Arroyo, I., Woolf, B., Barto, A., Fisher, D., and Mahadevan, S (2005). *Evaluating the Feasibility of Learning Student Models from Data*. American Association for Artificial Intelligence Workshop on Educational Data Mining.