Adaptive Testing using a General Diagnostic Model

Jill-Jênn Vie[1], Fabrice Popineau[1], Yolaine Bourda[1], and Éric Bruillard[2]

[1] LRI – Bât. 650 Ada Lovelace, Université Paris-Sud, 91405 Orsay, France,
{jjv, popineau, bourda}@lri.fr,
[2] ENS Cachan – Bât. Cournot, 61 avenue du Président Wilson, 94235 Cachan, France
eric.bruillard@ens-cachan.fr

Abstract. In online learning platforms such as MOOCs, computerized assessment needs to be optimized in order to prevent boredom and dropout of learners. Indeed, they should spend as little time as possible in tests and still receive valuable feedback. It is actually possible to reduce the number of questions for the same accuracy with computerized adaptive testing (CAT): asking the next question according to the past performance of the examinee. CAT algorithms are divided in two categories: summative CATs, that measure the level of examinees, and formative CATs, that provide feedback to the examinees at the end of the test by specifying which knowledge components need further work. In this paper, we formalize the problem of test-size reduction by predicting student performance, and propose a new hybrid CAT algorithm `GenMA` based on the general diagnostic model, that is both summative and formative. Using real datasets, we compare our model to popular CAT models and show that `GenMA` achieves better accuracy while using fewer questions than the existing models.

1 Introduction

Computerized assessments are becoming increasingly popular. Meanwhile, the Obama administration has urged schools to make exams less onerous and more purposeful [1]. To reduce over-testing, we need to optimize the time spent on tests, asking only informative questions about the learners' ability or knowledge. This is the idea behind Computerized Adaptive Testing (CAT): selecting the next question to ask according to the previous answers of the examinee. As an example, 238,536 such adaptive tests have been administered by the Graduate Management Admission Council in 2012–2013 [2] and adaptive assessment is getting more and more necessary in the current age of massive online open courses (MOOC), in order to minimize dropout.

Primarily, CATs have been relying on item response theory, that provides a framework to measure effectively scores called *latent traits* in order to rank students on a scale. The idea is to calibrate the difficulty of questions using a history of people having already taken the test. In 2001, The No Child Left Behind Act has called for more formative assessments, providing feedback to learners and

teachers at the end of the test. Such formative assessments may detect students with cognitive disabilities or simply build a profile that specifies which knowledge components seem to be mastered and which ones do not. A straightforward application would be a personal assistant that asks a few questions, then highlights the points that need further work, and possibly suggests useful material for remediation. In 2003, to address this need, new CATs have been developed relying on cognitive diagnosis models, the most popular being the DINA model [3] based on a q-matrix: a matrix that maps items (aka questions) to knowledge components involved in their resolution. Other cognitive models, less known, tend to unify scoring and formative assessments, but to date, they have not be used for adaptive testing [4].

In this paper, we formalize the problem of *test-size reduction by predicting student performance* (TeSR-PSP), inspired by [5] and present a new algorithm for CAT called `GenMA`, based on the general diagnostic model [6] that encompasses both the recovery of the latent knowledge components (KC) and for each KC, a degree of proficiency represented by a difficulty parameter.

To compare our algorithm for adaptive testing to the other ones mentioned above, we present an experimental protocol and execute it on real data. We show that `GenMA` outperforms existing models.

Our paper is organized as follows. In Section 1, we present the related work in CAT models. In Section 2, we formalize the problem of test-size reduction by predicting student performance and our new model, `GenMA`. In Section 3, we present an experimental protocol devised to compare these models, the real dataset used for evaluation, and our results. Finally, we discuss further work.

2 Related Work

2.1 Computerized Adaptive Testing and the problem of test-size reduction

In a non-adaptive test, every examinee receives the same set of questions. In an adaptive test, the next item asked by the system is chosen according to a certain criterion (the *selection item rule*), until the termination criterion holds, for example until a threshold over the parameters of the chosen model is guaranteed. Therefore, asking questions in an adaptive way means that the next question can be chosen as a function of the previous responses of the examinee.

In the problem of *test-size reduction* [5], one wants to reduce the number of questions asked as much as possible. Given a student model, we thus need to carefully choose the next question in order to still recover the model parameters. Formally, we want to decrease as much as possible the distance between the estimated and true user parameters after each question.

But in real data analysis, the true user parameters are unknown. For their evaluation, [5] replace the true user parameters with the estimated parameters they obtain after all questions have been asked, even if those estimated parameters do not fit the data at all.

In what follows, we will assume n learners take a test of a total of m questions. We assume the student data is dichotomous, which means every student either fails on succeeds over an item.

2.2 Item Response Theory (IRT)

The most simple model in item response theory for adaptive testing is the Rasch model, also known as 1-parameter logistic model. It models the behavior of a learner $i \in \{1, \ldots, n\}$ with a single parameter $\theta_i \in \mathbf{R}$ called ability, and models the item $j \in \{1, \ldots, m\}$ with a single parameter $d_j \in \mathbf{R}$ called difficulty. The tendency for a learner to solve an item only depends on the difference between the difficulty and the ability:

$$Pr(\text{“learner } i \text{ answers item } j\text{”}) = \Phi(\theta_i - d_j)$$

where $\Phi : x \mapsto 1/(1 + e^{-x})$ is the logistic function.

Being a unidimensional model, the Rasch model alone is not suitable for fine-grained feedback: it can only provide the level of the examinee at the end of the test. Still, it is really popular because of its simplicity, its stability and its sound mathematical framework [7, 8]. Also, [9] has showed that if the items are splitted into categories, the Rasch model is enough to provide to the examinee a useful deviation profile, specifying which category subscores were lower or higher than expected.

It is natural to extend the Rasch model to multidimensional abilities. In Multidimensional Item Response Theory (MIRT) [10], both learners and items are modelled by vectors of a certain dimension d, and the tendency for a learner to solve an item depends only on the dot product of those vectors. Thus, if learner $i \in \{1, \ldots, n\}$ is modelled by vector $\boldsymbol{\theta_i}$ and item $j \in \{1, \ldots, m\}$ is modelled by vector $\boldsymbol{d_j}$:

$$Pr(\text{“learner } i \text{ answers item } j\text{”}) = \Phi(\boldsymbol{\theta_i} \cdot \boldsymbol{d_j}).$$

Thus, a learner has greater chance to solve items correlated with its ability. Nevertheless, those richer models involve many more parameters, and have proven to be much harder to converge [7].

2.3 Cognitive Diagnosis

[11] have used adaptive testing strategies applied to cognitive diagnosis (CD) models, notably the DINA model. These cognitive models rely on a specification of the knowledge components (KCs) involved in the resolution of the items proposed in the test, in the form of a q-matrix, which simply maps items to KCs: q_{ik} is 1 if item i involves the KC k, 0 otherwise. Several algorithms have been proposed and compared for CD-CATs, using for example Kullback-Leibler divergence [3, 12].

If there are K KCs involved in a test, the learner can be modelled by a vector of K bits called *state*, specifying which KCs are mastered and which ones do not. Knowing the state of a learner, we can infer his performance over the different questions of the test. Slip and guess parameters capture careless errors. Throughout the assessment, a probability distribution over the 2^K states is maintained, and updated after each answer in order to fit the learner's behavior. In the particular case of the DINA model, the KCs involved in the resolution of an item are required to solve it. If the learner masters all KCs required, it still has a probability to slip over the question; if it lacks a KC, it still has a probability to guess correctly the answer.

3 Our Contribution

3.1 TeSR-PSP: Test-Size Reduction by Predicting Student Performance

In this paper, we propose a new problem called *test-size reduction by predicting student performance*: if we can ask only k questions in an adaptive way, which ones should we pick so as to predict the examinee's performance over the remaining questions of the test?

Usually, adaptive tests keep going until a suitable confidence interval over the learner parameters is obtained. In our case, we want to specify in advance the maximal number of questions that will be asked to every student, in order to prevent boredom from the learner.

3.2 `GenMA`: Using the General Diagnostic Model for Adaptive Testing

[6] has proposed a unified model that takes many existing IRT models and cognitive models as special cases: the general diagnostic model for partial credit data:

$$Pr(\text{“learner } i \text{ answers item } j\text{”}) = \Phi\left(\beta_i + \sum_{k=1}^{K} \theta_{ik} q_{jk} d_{jk}\right)$$

where K is the number of KCs involved in the test, β_i is the main ability of learner i, θ_{ik} its ability for KC k, q_{jk} is the (j,k) entry of the q-matrix which is 1 if KC k is involved in the resolution of item j, 0 otherwise, d_{jk} the difficulty of item j over KC k. Please note that this model is similar to the MIRT model specified above, but only parameters that correspond to a nonzero entry in the q-matrix are taken into account.

To the best of our knowledge, this model has not been used in adaptive testing [4]. This is what we present in this paper: `GenMA` relies on a general diagnostic model, thus requires the specification of a q-matrix by an expert. The parameters d_{jk} for every item j and KC k are calibrated using the history

of answers from a test and the Metropolis-Hastings Robbins-Monro algorithm [13, 14]. For the selection item rule of `GenMA`, we choose to maximize the Fisher information at each step, details of the implementation can be found in [13]. The problem TeSR-PSP becomes: after k questions asked to a certain learner i, how to estimate its main ability β_i and ability for each KC θ_{ik} that can explain its behavior throughout the test?

In real tests, items usually rely on only few KCs, hence there are fewer parameters to estimate than in a general MIRT model, which explains why the convergence is easy to obtain for `GenMA`. We can thus use the general diagnostic model to create an adaptive test that makes best of possible worlds: providing feedback under the form of degrees of proficiency over several KCs at the end of test, represented by the vector $\theta_i = (\theta_{i1}, \ldots, \theta_{iK})$, and being easy to converge. `GenMA` is both summative and formative, thus a hybrid model. Such feedback can be aggregated at various levels (e.g., from student, to class, to school, to city, to country) in order to enable decision-making [9, 15].

4 Evaluation

In this section, we detail the experimental protocol used to compare the following models for TeSR-PSP: the Rasch model, the DINA model and `GenMA`. For the sake of equality, we decide to define the same selection item rule for all models: all of them pick the question that maximizes Fisher information, which means the question of estimated probability closest to 1/2.

4.1 Experimental protocol

Our experimental protocol is based on double cross-validation. For each experiment, we need:

- a *train* student set, in order to calibrate the parameters of our model (for example, the difficulty of questions in the case of the Rasch model);
- a *test* student set, which will take our adaptive test;
- a *validation* question set V_Q, which is used for training, but kept out of the adaptive tests, used only to evaluate the prediction of performance of the students from the test set.

To evaluate the score of a model for our problem, we first train it using the train student set. Then, for each student from the test set, we let the model pick a question, we reveal the student answer for this question, the model updates its parameters accordingly and outputs a probability of correctly answering the questions from the validation set, that we can evaluate using negative log-likelihood, hereby denoted as *error*:

$$error(pred, truth) = \sum_{q \in V_Q} truth_q \log pred_q + (1 - truth_q) \log(1 - pred_q).$$

Then the model picks the second question, and so on. Thus, after k questions we can compute a prediction error over the validation question set for every model and every test student.

4.2 Real dataset: fraction subtraction

Tatsuoka's fraction subtraction dataset contains the dichotomous responses of 536 middle school students over 20 fraction subtraction test items. The corresponding q-matrix maps the 20 items to the following 8 knowledge components (KCs):

- convert a whole number to a fraction,
- separate a whole number from a fraction,
- simplify before subtracting,
- find a common denominator,
- borrow from whole number part,
- column borrow to subtract the second numerator from the first,
- subtract numerators,
- reduce answers to simplest form.

The cross-validation was performed using a random split into training and test sets: the split was 5-fold over the students and 4-fold over the questions: each student set had a size of 20% while each validation question set had a size of 25% so 5 questions, therefore there were 20 experiments in total, of which the mean error was computed.

4.3 Implementation details

Our Rasch model implementation comes from the `ltm` R package. We made our custom implementation of the DINA model but the slip and guess calibration is held by the `CDM` package. `GenMA` is built upon the `mirt` package [13].

4.4 Results

For each number of questions asked from 1 to 15, we plotted the mean error of each model (Rasch model, DINA model and `GenMA`) over the test student set in Figure 1, and as an insight easier to comprehend, the mean number of questions incorrectly guessed in Figure 2.

Figure 1 shows that 8 questions over 15 are enough for the Rasch model to converge on the fraction subtraction dataset. Figure 2 shows that no matter how many questions are asked, the Rasch and DINA models can't predict correctly more than 4 questions in average over 5 in the validation question set, while `GenMA` can achieve this accuracy with only 4 questions, then keeps on improving its predictions. The DINA model takes a long time to converge because the first questions require a single KC, therefore they do not bring a lot of information about the user state. But still, the simplest, unidimensional Rasch model performs surprisingly well compared to `GenMA` which is over 8 dimensions, one per KC of the q-matrix.

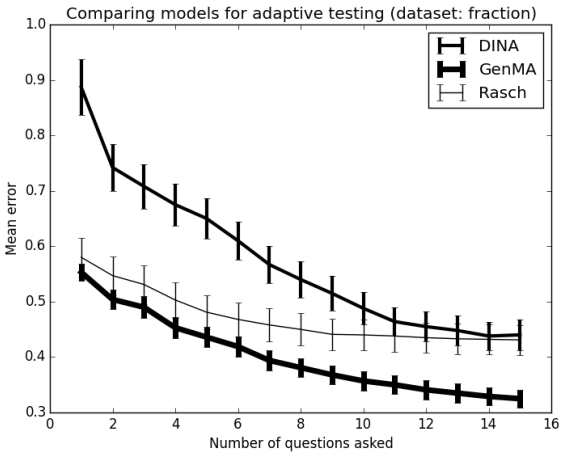

Fig. 1. Comparing adaptive testing models. Evolution of error (negative log-likelihood) over the validation question set, after a certain number of questions have been asked.

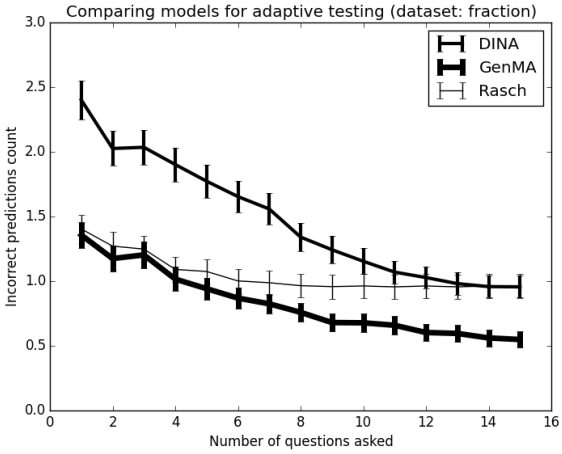

Fig. 2. Comparing adaptive testing models. Evolution of the number of questions in the validation question set incorrectly predicted, after a certain number of questions have been asked.

5 Conclusion and Future Work

In this paper, we formulated the problem of test-size reduction by predicting student performance, and presented our new adaptive testing algorithm `GenMA` to tackle it, based on the general diagnostic model. As this model is richer than other models as Rasch or DINA, it could be prone to overfitting: having more parameters, it may have a good score over train data but poor generalization over the test data. But we showed it actually achieves a better accuracy at predicting student performance, using fewer questions than the existing models on a real dataset.

The idea of a hybrid model combining several KCs and weights for each of them is not new: MIRT models can be seen this way, but there are many parameters to estimate, leading to convergence issues. [5] presented sparse factor analysis (SPARFA), a model that combine q-matrices and weights but their KCs are specified automatically, not by experts, thus it is not possible to provide a feedback at the end of the test.

In order to overcome the complexity of $O(2^K)$ of the DINA model, some knowledge representations such as Attribute Hierarchy Model [16, 17] or Knowledge Space Theory [18, 19] have been devised, relying on dependencies over KCs in the form of a directed acyclic graph. We would like to compare these approaches to `GenMA`. Also, richer models using ontologies [20, 21] are a promising direction of research.

6 Acknowledgements

This work is supported by the Paris-Saclay Institut de la Société Numérique funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02.

References

1. Zernike, K.: Obama administration calls for limits on testing in schools, (2015).
2. Council, G.M.A.: Profile of GMAT Candidates – Executive Summary, (2013).
3. Huebner, A.: An overview of recent developments in cognitive diagnostic computer adaptive assessments. Practical Assessment, Research & Evaluation. 15, n3 (2010).
4. Yan, D., Davier, A.A. von, Lewis, C.: Computerized multistage testing. (2014).
5. Lan, A.S., Waters, A.E., Studer, C., Baraniuk, R.G.: Sparse factor analysis for learning and content analytics. The Journal of Machine Learning Research. 15, 1959–2008 (2014).
6. Davier, M.: A general diagnostic model applied to language testing data. ETS Research Report Series. 2005, i–35 (2005).
7. Desmarais, M.C., Baker, R.S.: A review of recent advances in learner and skill modeling in intelligent learning environments. User Modeling and User-Adapted Interaction. 22, 9–38 (2012).
8. Bergner, Y., Droschler, S., Kortemeyer, G., Rayyan, S., Seaton, D., Pritchard, D.E.: Model-based collaborative filtering analysis of student response data: Machine-learning item response theory. International Educational Data Mining Society. (2012).

9. Verhelst, N.D.: Profile analysis: A closer look at the PISA 2000 reading data. Scandinavian Journal of Educational Research. 56, 315–332 (2012).
10. Reckase, M.: Multidimensional item response theory. Springer (2009).
11. Xu, X., Chang, H., Douglas, J.: A simulation study to compare CAT strategies for cognitive diagnosis. In: Annual meeting of the american educational research association, chicago (2003).
12. Cheng, Y.: When cognitive diagnosis meets computerized adaptive testing: CD-CAT. Psychometrika. 74, 619–632 (2009).
13. Chalmers, R.P.: mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software. 48, 1–29 (2012).
14. Cai, L.: Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. Journal of Educational and Behavioral Statistics. 35, 307–335 (2010).
15. Shute, V., Leighton, J.P., Jang, E.E., Chu, M.-W.: Advances in the science of assessment. Educational Assessment. (2015).
16. Leighton, J.P., Gierl, M.J., Hunka, S.M.: The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. Journal of Educational Measurement. 41, 205–237 (2004).
17. Rupp, A., Levy, R., Dicerbo, K.E., Sweet, S.J., Crawford, A.V., Calico, T., Benson, M., Fay, D., Kunze, K.L., Mislevy, R.J., others: Putting ECD into practice: The interplay of theory and data in evidence models within a digital learning environment. JEDM-Journal of Educational Data Mining. 4, 49–110 (2012).
18. Doignon, J.-P., Falmagne, J.-C.: Knowledge spaces. Springer Science & Business Media (2012).
19. Lynch, D., Howlin, C.P.: Real world usage of an adaptive testing algorithm to uncover latent knowledge, (2014).
20. Mandin, S., Guin, N.: Basing learner modelling on an ontology of knowledge and skills. In: Advanced learning technologies (iCALT), 2014 iEEE 14th international conference on. pp. 321–323. IEEE (2014).
21. Kickmeier-Rust, M.D., Albert, D.: Competence-based knowledge space theory. Measuring and Visualizing Learning in the Information-Rich Classroom. 109 (2015).