A Review of Recent Advances in Adaptive Assessment

Jill-Jênn Vie, Fabrice Popineau, Éric Bruillard, Yolaine Bourda

Chapter 4
A Review of Recent Advances in Adaptive Assessment

Jill-Jênn Vie, Fabrice Popineau, Éric Bruillard, and Yolaine Bourda

Jill-Jênn Vie, Fabrice Popineau, Yolaine Bourda (+33 169851480)
LRI – Bât. 650 Ada Lovelace, Université Paris-Sud, 91405 Orsay, France
{jjv, fabrice.popineau, yolaine.bourda}@lri.fr JJV (+33 642623974) FP (+33 169851950)

Éric Bruillard
ENS Cachan – Bât. Cournot, 61 avenue du Président Wilson, 94235 Cachan, France
eric.bruillard@ens-cachan.fr (+33 147402457)

Abstract Computerized assessments are an increasingly popular way to evaluate students. They need to be optimized so that students can receive an accurate evaluation in as little time as possible. Such optimization is possible through learning analytics and computerized adaptive tests (CATs): the next question is then chosen according to the previous responses of the student, thereby making assessment more efficient. Using the data collected from previous students in non-adaptive tests, it is thus possible to provide formative adaptive tests to new students by telling them what to do next. This chapter reviews several models of CATs found in various fields, together with their main characteristics. We then compare these models empirically on real data. We conclude with a discussion of future research directions for computerized assessments.

Keywords: latent knowledge extraction, item response theory, q-matrix, cognitive diagnosis models, adaptive testing, knowledge space theory

List of Abbreviations

CAT	Computerized adaptive testing
CD	Cognitive diagnosis
DINA	Deterministic input, noisy and
ECPE	Examination for the Certificate of Proficiency in English
KC	Knowledge components
GenMA	General Multidimensional Adaptive
GMAT	Graduate Management Admission Test
GRE	Graduate Record Examination
LA	Learning analytics
MIRT	Multidimensional item response theory
MOOC	Massive online open course
MST	Multistage testing

PCA	Principal components analysis
TIMSS	Trends in International Mathematics and Science Study

4.1 Introduction

Today, educational assessments are often automatized, so we can store and analyze student data in order to provide more accurate and shorter tests for future learners. The *learning analytics* process consists in collecting data about learners, discovering hidden patterns that can lead to a more effective learning experience, and constantly refining models using new learner data (Chatti et al. 2012). Learning analytics for adaptive assessment have specific and well-defined objectives: they must improve the efficiency and effectiveness of the learning process, and tell learners what to do next by adaptively organizing instructional activities (Chatti et al. 2012). Reducing the test length in needed even more as students today are over-tested (Zernike 2015), leaving less time for instruction.

Traditionally, models used for adaptive assessment have been mostly *summative*: they measure or rank effectively examinees, but do not provide any other feedback. This is in particular the case of models encountered in item response theory (Hambleton and Swaminathan 1985). Recent advances have focused on *formative assessments* (Ferguson 2012, Huebner 2010), providing more useful feedback for both the learner and the teacher; hence, they are more useful to the learning analytics community. Indeed, Tempelaar et al. (2015) have shown that computer-assisted formative assessments have high predictive power for detecting underperforming students and estimating academic performance.

In this chapter, we prove that such adaptive strategies can be applied to formative assessments, in order to make tests shorter and more useful. Our primary focus is the assessment of knowledge and we do not consider dimensions of conscientiousness, i.e., perseverance, organization, carefulness, responsibility. Our second focus is to provide useful feedback at the end of the test. Such feedback can be aggregated at various levels (e.g., at the level of an individual student, of a class, or of a school, district, state, or country) for decision-making purposes (Shute et al. 2016).

We assume that data is provided as *dichotomous response patterns*, i.e., learners answer each question either correctly or incorrectly. A general method is to train user models so they can help uncover the latent knowledge of new examinees using fewer, carefully chosen questions. We here develop a framework that relies solely on dichotomous data in order to compare different adaptive models on the same data. Our approach is thus generic and can be specialized for different environments, e.g., serious games. Based on our analysis, one can choose the best model suitable to their individual needs.

This chapter is organized as follows. First in Section 4.2, we present the learning analytics methods that will be used in the chapter. Then in Section 4.3, we describe the adaptive assessment models used in diverse fields, ranging from psychometrics to machine learning. Later in Section 4.4, we present a protocol to compare adaptive

assessment strategies for predicting student performance, and expose our experimental results on real data in Section 4.5. In Section 4.6, we highlight which models suit which use cases, specify the limitations of our approach in Section 4.7, discuss possible directions for the future of assessment in Section 4.8 and finally draw our conclusions in Section 4.9.

4.2 Learning Analytics

Educational data mining and learning analytics are two research communities that analyze educational data, typically collected in online environments and platforms. The former focuses on automated adaptation while the latter provides tools for human intervention. Indeed, various dashboards, visualizations and analytics packages can help inform pedagogical decision-making. Faculty, instructional designers, and student support services often use data to improve teaching, learning, and course design.

Among the objectives of learning analytics (LA), Chatti et al. (2012) describe the need for intelligent feedback in assessment, and the problem of choosing the next activity to present to the learner. To address these needs, they highlighted the following classes of methods: statistics, information visualization, data mining and social network analysis. In this chapter, we describe the methods used to provide adaptive assessments.

Adaptive assessments can lead to improved personalization, by organizing learning resources. For example, the problem of *curriculum sequencing* studies how we can choose learning paths in a space of learning objectives (Desmarais and Baker 2012). It aims to use skills assessment to tailor the learning content, based on as little evidence as possible. As stated by Desmarais and Baker (2012), "The ratio of the amount of the evidence to the breadth of the assessment is particularly critical for systems that cover a large array of skills, as it would be unacceptable to ask hours of questions before making a usable assessment."

In educational systems, there is an important difference between *adaptivity*, the ability to modify course materials using different parameters and a set of pre-defined rules, and *adaptability*, the possibility for learners to personalize the course materials by themselves. As Chatti et al. (2012) indicate, "more recent literature in personalized adaptive learning have criticized that traditional approaches are very much top-down and ignore the crucial role of the learners in the learning process." There should be a better balance between giving learners what they need to learn (i.e. adaptivity) and giving them what they want to learn (i.e. adaptability), the way they want to learn it (e.g., giving them more examples, or more exercises, depending on what they prefer). In either case, learner profiling is a crucial task.

As a use case scenario, let us consider users who register on a massive online open course (MOOC). As these users may have acquired knowledge from diverse backgrounds, some may be missing some prerequisites of the course, whereas other could afford to skip some chapters of the course. Therefore, it would be useful to

adaptively assess user needs and preferences, to filter the content of the course accordingly and minimize information overload. Lynch and Howlin (2014) describe such an algorithm to uncover the latent knowledge state of a learner, by asking a few questions at the beginning of the course. Another lesser-known use case is the automated generation of testlets of exercises on demand, that reduce the costs of practice testing.

In learning analytics, methods in data mining include machine learning techniques such as regression trees for prediction. For instance, gradient boosting trees can be used to highlight which variables are the most informative to explain why a MOOC user obtained a certificate (or failed to obtain it). Gradient boosting trees have also been successful to tackle prediction problems, notably in data science challenges, because they can integrate heterogeneous values (categorical variables and numerical variables) and they are robust to outliers. It is surprising to see that learning analytics methods produced so many models to predict some objective from a fixed set of variables, and so few models to assess the learner about their needs and preferences. We believe that a lot of research can still be done towards more interactive models in learning analytics.

Recommender systems are another tool to aggregate data about users in order to recommend relevant resources (such as movies, products). They are increasingly used in technology-enhanced learning research as a core objective of learning analytics (Chatti et al. 2012, Manouselis et al. 2011, Verbert et al. 2011). Most recommender systems rely on *collaborative filtering*, a method that makes automated predictions about the interests of a user, based on information collected from many users. The intuition is that a user may like items that similar users have liked in the past. In our case, a learner may face difficulties similar to the ones faced by learners with similar response patterns. There are open research questions on how algorithms and methods have to be adapted from the field of commercial recommendations. Still, we believe that existing techniques can be applied to adaptive assessment.

Another approach, studied in cognitive psychology, is to measure the *response time* during an assessment. Indeed, the amount of time needed by a person needs to answer a question can give some clues about the cognitive process. To do so, sophisticated statistical models are needed (Chang, 2014); we do not consider them in this chapter.

4.3 Adaptive Assessments

Our goal is to filter the questions to ask to a learner. Instead of asking the same questions to everyone, the so-called computer adaptive tests (CATs) (van der Linden and Glas 2010) select the next question to ask based on the previous answers, thus allowing adaptivity at each step. The design of CATs relies on two criteria: a *termination criterion* (e.g., a number of questions to ask), and a *next item criterion*. While the termination criterion is not satisfied, questions are asked according to the next item criterion, which picks questions, e.g., that are the most informative about

the learner's ability or knowledge. Lan et al. (2014) have proven that such adaptive tests needed fewer questions than non-adaptive tests to reach the same prediction accuracy.

This gain in performance is important: shorter tests are better for the system, because they reduce load, and they are better for the learner, who may be frustrated or bored if they need to give too many answers (Lynch and Howlin 2014, Chen et al. 2015). Thus, adaptive assessment is more and more useful in the current age of MOOCs, where motivation plays an important role (Lynch and Howlin 2014). In real-life scenarios, however, more constraints need to be taken into account. First, the computation of criteria should be done in a reasonable time; hence the time complexity of the approaches is important. Second, assessing skills must be performed under uncertainty: a learner may *slip*, i.e., accidentally or carelessly fail an item that they could have solved, or they may *guess*, i.e., correctly answer an item by chance. This is why adaptive assessment cannot simply perform a binary search over the ability of the learner, i.e., asking a more difficult question if they succeed and an easier question if they fail. Thus, we need to use more robust methods, such as probabilistic models for skill assessment.

CATs have been extensively studied over the past years, and they have been put into practice. For instance, the Graduate Management Admission Council has administered 238,536 adaptive tests of this kind in 2012–2013 through the Graduate Management Admission Test (GMAT) (Graduate Management Admission Council 2013). Given a student model (Peña-Ayala 2014), the objective is to provide an accurate measurement of the parameters of an upcoming student while minimizing the number of questions asked. This problem has been referred to as *test-size reduction* (Lan et al. 2014), and it is also related to predicting student performance (Bergner et al. 2012, Thai-Nghe et al. 2011). In machine learning, this approach is known as *active learning*: adaptively query the informative labels of a training set in order to optimize learning.

Several models can be used, depending on the purpose of the assessment, e.g., estimating a general level of proficiency, providing diagnostic information, or characterizing knowledge (Mislevy et al. 2012). At the end of the test, rich feedback can help teachers identify at-risk students. It also protects against perseveration errors when students respond incorrectly on a practice test (Dunlosky et al. 2013). In what follows, we describe those models under the following categories: item response theory for summative assessment (Section 4.3.1), cognitive models for formative assessment (Section 4.3.2), more complex knowledge structures (Section 4.3.3), adaptive assessment and recommender systems (Section 4.3.4), exploration and exploitation trade-off (Section 4.3.5), and multistage testing (Section 4.3.6).

4.3.1 Psychometrics: Measuring Proficiency using Item Response Theory

The simplest model for adaptive testing is the *Rasch model*, also known as the 1-parameter logistic model: it falls into the data mining category of LA. This model represents the behavior of a learner with a single latent trait, called *ability*, and the items or tasks with a single parameter, called *difficulty*. The tendency for a learner to solve a task only depends on the difference between the difficulty of the task and the ability of the learner. Thus, if a learner i has ability θ_i and wants to solve an item j of difficulty d_j, the probability that the learner i answers the item j correctly is given by Equation (4.1), where $\Phi : x \mapsto 1/(1 + e^{-x})$ is the *logistic function*:

$$Pr(\text{“learner } i \text{ answers item j”}) = \Phi(\theta_i - d_j). \tag{4.1}$$

Of course, we cannot specify all difficulty values by hand, as it would be time-consuming and probably inaccurate (i.e., be too subjective, and poorly fit student data). Fortunately, the Rasch model makes it possible to estimate parameters efficiently: using former student data, we can calibrate item difficulties and learner abilities automatically, computing the maximum likelihood estimates. In particular, this estimation process does not depend on any domain knowledge.

When a new user takes a test, the observed variables are its outcomes over the questions that are asked to the user, and the hidden variable we want to estimate is the ability of the user, given the known difficulty parameters. This is usually performed using maximum likelihood estimation: we can easily do this computationally, using Newton's method to find the zeroes of the derivative of the likelihood function. Therefore, the adaptive process can be phrased as follows: given an estimate of the learner's ability, which question outcome will be the most useful to refine this estimate? Indeed, we can quantify the information that each item j provides over the ability parameter: this can be done using *Fisher information*, defined as the variance of the gradient of the log-likelihood with respect to the ability parameter, given by Equation (4.2):

$$I_j(\theta_i) = E\left(\left(\frac{\partial}{\partial\theta}\log f(X_j, \theta_i)\right)^2 \middle| \theta_i\right). \tag{4.2}$$

where X_j is the binary outcome of the learner i over the item j and $f(X_j, \theta_i)$ is the probability function for X_j depending on $\theta_i : f(X_j, \theta_i) = \Phi(\theta_i - d_j)$.

Therefore, an adaptive assessment can be designed as follows: given the learner's current ability estimate, pick the question which yields the most information about the ability, update the estimate according to the outcome (i.e., whether the user answered correctly or incorrectly), and so on. At the end of the test, one can visualize the whole process like in Fig. 4.1. As we can see, the confidence interval for the ability estimate is refined after each outcome.

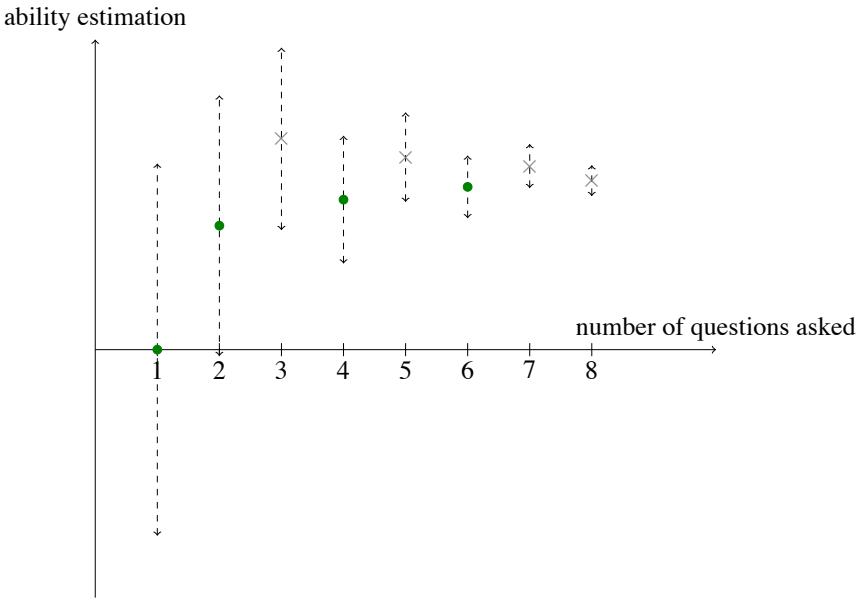

Fig. 4.1 Evolution of the ability estimate throughout an adaptive test based on the Rasch model. Filled circles denote correct answers while crosses denote incorrect answers.

As the Rasch model is a unidimensional model, it is not suitable for cognitive diagnosis. Still, it is really popular because of its simplicity, its stability, and its sound mathematical framework (Desmarais and Baker 2012, Bergner et al. 2012). Also, Verhelst (2012) has showed that, if the items are split into categories, we can provide to the examinee a useful deviation profile, specifying the categories where the subscores were higher or lower than expected. Specifically, let us consider that, in each category, an answer gives one point if correct, and no point otherwise. The *subscores* are then the number of points obtained by the learner in each category, which sum up to the total score. Given the total score, we can then compute the expected subscore of each category by simply using the Rasch model. Finally, the *deviation profile*, namely, the difference between the observed and expected subscores, provides a nice visualization of the categories that need further work: see Fig. 4.2 for an example. Such deviation profiles can be aggregated across a country to highlight the strong and weak points of its students, which can help identify deficiencies in the national curriculum. These profiled can then be compared worldwide in studies of international assessments, such as the Trends in International Mathematics and Science Study (TIMSS). For instance, Fig. 4.2 presents the TIMSS 2011 dataset of proficiency in mathematics, highlighting the fact that Romania is stronger in Algebra than expected, while Norway is weaker in Algebra than expected. This belongs to the information visualization class of learning analytics methods, and shows what can be done using the simplest psychometric model and the student data only.

In adaptive testing, however, we do not observe all student responses, but only the answers to the subset of questions that we asked, and these may differ from a student to another. It is still possible to compute the deviation profile within this subset, but it cannot be aggregated to a higher level in this fashion, because of the bias induced by the adaptive process.

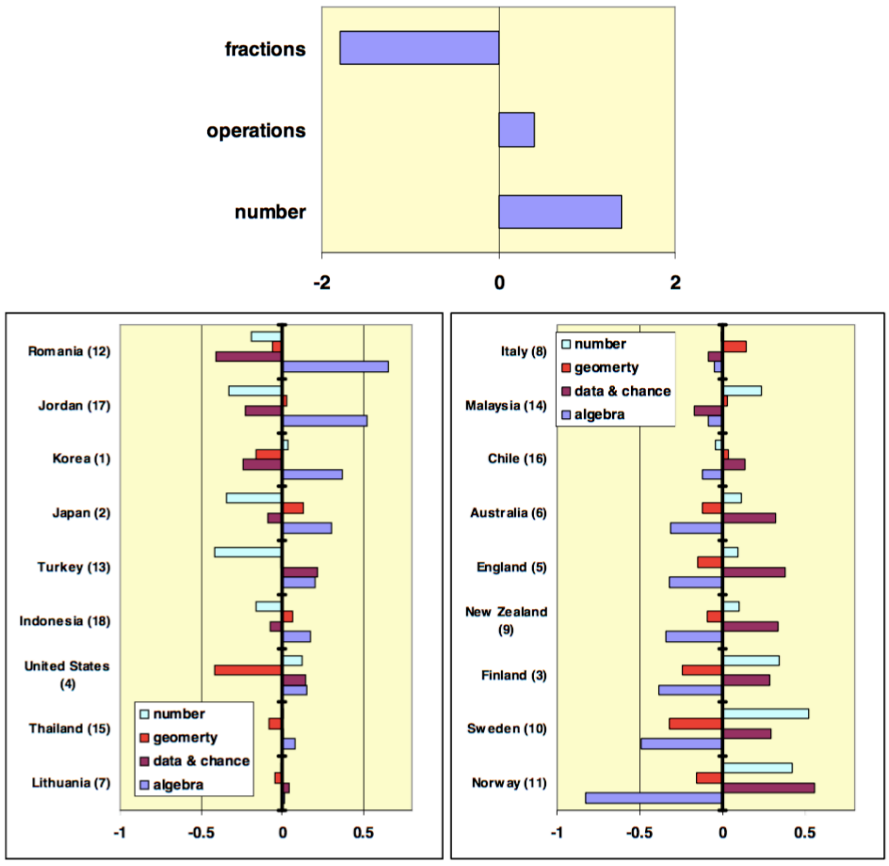

Fig. 4.2 Above, the deviation profile of a single learner. Below, the deviation profile of different countries on the TIMSS 2011 math dataset, from the presentation of N.D. Verhelst at the workshop Psychoco 2016.

A natural direction to extend the Rasch model is to study multidimensional abilities. In Multidimensional Item Response Theory (MIRT) (Reckase 2009), both learners and items are modeled 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, a learner has a greater chance to solve items that are correlated with their ability vector, and asking a question brings information in the direction of its item vector.

Thus, if learner $i \in \{1, \dots, n\}$ is modelled by vector $\boldsymbol{\theta_i} \in \mathbf{R}^d$ and item $j \in \{1, \dots, m\}$ is modelled by vector $\boldsymbol{d_j} \in \mathbf{R}^d$, the probability that the learner i answers the item j correctly is given by Equation (4.3):

$$Pr(\text{“learner } i \text{ answers item } j\text{”}) = \Phi(\boldsymbol{\theta_i} \cdot \boldsymbol{d_j}). \tag{4.3}$$

Using this model, the Fisher information becomes a matrix. When trying to ask the most informative questions, we may either choose to maximize the determinant of this matrix (“D-rule”), or choose to maximize the trace (“T-rule”). The D-rule chooses the item that provides the maximum volume of information, and hence the largest reduction of volume in the variance of the ability estimate. By contrast, the T-rule chooses an item that attempts to increase the average information about each component of the ability, ignoring the covariance between components.

MIRT can be restated as a matrix factorization problem, given by Equation (4.4):

$$M \simeq \Phi(\Theta D^T) \tag{4.4}$$

where M is the $n \times m$ student data, Θ is the $n \times r$ *learner matrix* composed of the vectors of all learners, and D is the $m \times r$ *item matrix* which contains of all item vectors.

Nevertheless, those richer models involve many more parameters: d parameters are estimated for each of the n learners, and d parameters are estimated for each of the m items. Thus, this model is usually much harder to calibrate (Desmarais and Baker 2012, Lan et al. 2014).

4.3.2 Cognitive Diagnosis: Adaptive Assessment with Feedback

In cognitive diagnosis models, we assume that we can explain a student's success or failure on a learning task, based on whether they master (or fail to master) some *knowledge components* (KC). The point of these knowledge components is that they allow a transfer of evidence from one item to another. For instance, to evaluate correctly the sum $1/7 + 8/9$, a learner needs to know how to add numbers, and how to convert two fractions to the same denominator. By contrast, a learner that solves $1/7 + 8/7$ only needs to know how to add. To use these cognitive models, we need to specify, for each item proposed in the test, which KCs are required to solve it: this information is represented as a binary matrix, called the *q-matrix*. The q-matrix simply maps items to KCs: it is a transfer model. See Fig. 4.3 for a real-world example of a q-matrix.

	Knowledge components							
	1	2	3	4	5	6	7	8
Item 1	0	0	0	1	0	1	1	0
Item 2	0	0	0	1	0	0	1	0
Item 3	0	0	0	1	0	0	1	0
Item 4	0	1	1	0	1	0	1	0
Item 5	0	1	0	1	0	0	1	1
Item 6	0	0	0	0	0	0	1	0
Item 7	1	1	0	0	0	0	1	0
Item 8	0	0	0	0	0	0	1	0
Item 9	0	1	0	0	0	0	0	0
Item 10	0	1	0	0	1	0	1	1
Item 11	0	1	0	0	1	0	1	0
Item 12	0	0	0	0	0	0	1	1
Item 13	0	1	0	1	1	0	1	0
Item 14	0	1	0	0	0	0	1	0
Item 15	1	0	0	0	0	0	1	0
Item 16	0	1	0	0	0	0	1	0
Item 17	0	1	0	0	1	0	1	0
Item 18	0	1	0	0	1	1	1	0
Item 19	1	1	1	0	1	0	1	0
Item 20	0	1	1	0	1	0	1	0

Description of knowledge components:

1. convert a whole number to a fraction
2. separate a whole number from a fraction
3. simplify before subtracting
4. find a common denominator
5. borrow from whole number part
6. column borrow to subtract the second numerator from the first
7. subtract numerators
8. reduce answers to simplest form

Fig. 4.3 The q-matrix corresponding to Tatsuoka's (1984) fraction subtraction data set of 536 middle school students over 20 fraction subtraction test items. The matrix has 8 knowledge components, which are described on the right.

The DINA model ("Deterministic Input, Noisy And") assumes that the learner will solve a certain item i with probability $1 - s_i$ if they master every required KC, and will solve it with probability g_i otherwise. The parameter g_i is called the *guess parameter* of item i, and it represents the probability of guessing the right answer to item i without being able to solve it. The parameter s_i is called the *slip parameter* of item i: it represents the probability of slipping on item i, i.e., failing to answer it even when the correct KCs are mastered. By contrast, in the DINO model ("Deterministic Input, Noisy Or"), the learner solves an item with probability $1 - s_i$ whenever it masters one of the KCs for this item; if the learner masters none of them, the probability of solving the item is g_i.

The latent state of a learner is represented by a vector of K bits $(c_1, \ldots, c_K)$ where K is the total number of KCs. The vector indicates which KCs are mastered: for each KC k, the bit c_k is 1 if the learner masters the k-th KC, and 0 otherwise. Each time the learner answers an item, we obtain more information about their probable latent state. Xu et al. (2003) have used adaptive testing strategies in order to infer the latent state of the learner using few questions: this is called *cognitive diagnosis*

computerized adaptive testing (CD-CAT). Knowing the mental state of a learner, we can infer their behavior over the remaining questions in the test; we can then use this information to choose which questions to ask, as we will now describe. At each point in time, the system keeps a probability distribution over the 2^K possible latent states: this distribution is refined after each question, using Bayes' rule. A usual measure of uncertainty on the distribution is *entropy*, defined by Equation (4.5):

$$H(\mu) = -\sum_{c\in\{0,1\}^k} \mu(c) \log \mu(c). \tag{4.5}$$

Hence, to converge quickly into the true latent state, the best item to ask is the one that reduces average entropy the most (Doignon et Falmagne 2012, Huebner 2010). Other criteria have been proposed: for instance, we can ask the question that maximizes the *Kullback-Leibler divergence*, which measures the difference between two probability distributions (Cheng 2009). It is given by Equation (4.6):

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}. \tag{4.6}$$

As Chang (2014) states, "A survey conducted in Zhengzhou found that CD-CAT encourages critical thinking, making students more independent in problem solving, and offers easy to follow individualized remedy, making learning more interesting."

For large values of K, it may be intractable to maintain a probability distribution over the 2^K states. Hence, in practice, we often take $K \leq 10$ (Su et al. 2013). We can also reduce the complexity by assuming prerequisites between KCs: if mastering a KC implies that the student must master another KC, the number of possible states decreases, and so does the complexity. This approach is called the *Attribute Hierarchy Model* (Leighton et al. 2004): it can be used to represent knowledge more accurately and fit the data better (Rupp et al. 2012).

The q-matrix may be costly to build. Thus, devising a q-matrix automatically has been an open field of research. Barnes (2005) used a hill-climbing technique while Winters et al. (2005) and Desmarais et al. (2011) tried non-negative matrix factorization techniques to recover q-matrices from real and simulated multidisciplinary assessment data. Experimentally, these approaches can efficiently separate items in categories when the topics are clearly separated, e.g., French and Mathematics. Formally, *non-negative matrix factorization* tries to devise matrices with non-negative coefficients W and Q such that the original matrix M verifies $M \simeq WQ^T$. Additional constraints can be made: for instance, *sparse PCA* (Zou et al. 2006) looks for a factorization of the form $M \simeq WQ^T$ where Q is sparse, following the assumption that only few knowledge components are required for any one task. On the datasets we described in Section 4.5, the expert-specified q-matrix fitted the data better than a q-matrix devised automatically using sparse PCA. Further, even if we could fit the data better with an automatically devised q-matrix, this would not allow us to deduce human-readable names for the knowledge components. Lan et al. (2014) tried to circumvent this issue, by studying how to interpret a posteriori the columns of a q-matrix devised by an algorithm, with the help of expert-specified tags. A

more recent work from (Koedinger et al. 2012) managed to combine q-matrices from several experts using crowdsourcing in order to find better cognitive models that are still understandable for humans.

A natural goal is then to design a model that combines the best of both worlds, and represent which knowledge components are required for tasks, as well as some notion of the difficulty of tasks. Unified models have been designed towards this end, such as the *general diagnostic model for partial credit data* (Davier 2005), which generalizes both MIRT and some other cognitive models. It is given by Equation (4.7):

$$Pr(\text{“learner } i \text{ answers item } j\text{”}) = \Phi\left(\beta_i + \sum_{k=1}^{K} \theta_{ik}\, q_{jk} d_{jk}\right) \qquad (4.7)$$

where K is the total number of KCs involved in the test, β_i is the main ability of learner i, θ_{ik} is its ability for KC k, d_{jk} the difficulty of item j over KC k, and q_{jk} is the (j,k) entry of the q-matrix: 1 if KC k is involved in the resolution of item j, and 0 otherwise. Intuitively, this model is similar to the MIRT model presented above, but the dot product is computed only on part of the components. In other words, we consider a MIRT model where the number of dimensions is the number of KCs of the q-matrix : $d = K$. When we calibrate the feature vector of dimension d of an item, only the components that correspond to KCs involved in the resolution of this item are taken into account: see Fig. 4.4. This model has one important advantage: as few KCs are usually required to solve each item, this allows the MIRT parameter estimation to converge faster. Vie et al. (2016) used this model in adaptive assessment under the name GenMA (for General Multidimensional Adaptive). Another advantage of this model is that, at any point in the test, the ability estimate represents degrees of proficiency for each knowledge component. The GenMA model is therefore a hybrid model that combines the Rasch model and a cognitive model.

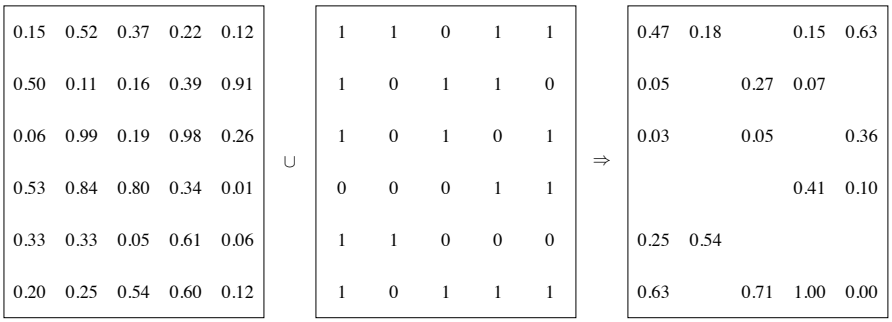

Fig. 4.4 The GenMA hybrid model, combining item response theory and a q-matrix.

4.3.3 Competence-based Knowledge Space Theory and Applications

Doignon et Falmagne (2012) have developed *knowledge space theory*, an abstract theory that relies on a partial order between subsets of a discrete knowledge space. Formally, let us assume that there is a certain number of KCs to learn, following a dependency graph specifying which KCs needs to be mastered before learning a certain KC. We present an example of dependency graph in Fig. 4.5. From this graph, one can compute the feasible knowledge states, i.e., the KCs that are actually mastered by the learner. For example, $\{a, b\}$ is a feasible knowledge state while the singleton $\{b\}$ is not, because a needs to be mastered before b. Thus, for this example there are 10 feasible knowledge states: $\emptyset$, $\{a\}$, $\{b\}$, $\{a, b\}$, $\{a, c\}$, $\{a, b, c\}$, $\{a, b, c, d\}$, $\{a, b, c, e\}$, $\{a, b, c, d, e\}$, $\{a, b, c, d, e, f\}$. An adaptive assessment can then uncover the knowledge state of the examinee, in a similar fashion to the Attribute Hierarchy Model described above at Section 4.3.2. Once the knowledge subset of a learner has been identified, this model can suggest to him the next knowledge components to learn in order to help them progress, through a so-called *learning path*. For instance, from the knowledge state $\{a\}$ on Fig. 4.5, the learner can choose whether to learn the KC b or the KC c first.

Falmagne et al. (2006) provide an adaptive test in order to guess effectively the knowledge space using entropy minimization, which is however not robust to careless errors. This model has been implemented in practice in the ALEKS system, which is used by millions of users today (Kickmeier-Rust and Albert 2015, Desmarais and Baker 2012).

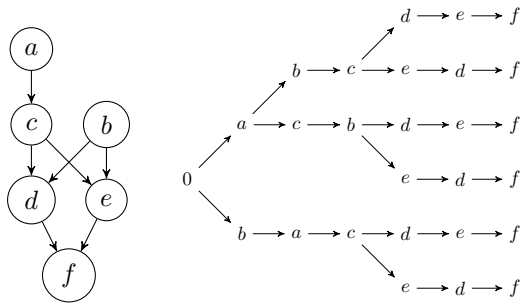

Fig. 4.5 On the left, an example of precedence diagram. On the right, the corresponding learning paths.

Lynch and Howlin (2014) have implemented a similar adaptive pretest at the beginning of a MOOC, in order to guess what the learner already masters, and help them jump directly to useful materials in the course. To address slip and guess parameters, they combine models from knowledge space theory and item response theory.

Another line of work has developed more fine-grained models for adaptive testing, by considering even richer domain representations such as an ontology (Mandin and Guin 2014, Kickmeier-Rust and Albert 2015) of the domain covered by the test. However, such knowledge representations are costly to develop.

4.3.4 Adaptive assessment and recommender systems

We now describe how two well-known problems from recommender systems find their counterparts in adaptive assessment. Recommender systems can recommend new items to a user based on their preferences on other items. Two approaches are used:

- *content-based recommendations*, that analyze the content of the items in order to devise a measure of similarity between items;
- *collaborative filtering*, where the similarity between items depends solely on user preferences, i.e., items that are liked by the same people are considered to be similar.

Overall, the aim of these approaches is to predict the preference of a user over an unseen item, based on their preferences over a fraction of the items that they know. In our case, we want to predict the performance of a user over a question that we did not ask yet, based on the previous performance of the user. Collaborative filtering techniques have been applied on student data in an user-to-resource fashion (Manouselis et al. 2011, Verbert et al. 2011) and in an user-to-task fashion (Toscher and Jahrer 2010, Thai-Nghe et al. 2011, Bergner et al. 2012).

All recommender systems face the *user cold-start problem*: given a new user, how to quickly recommend new relevant items to them? In technology-enhanced learning, the problem becomes: given a new learner, how to quickly identify the resources that they will need? To the best of our knowledge, the only work that references the cold-start problem in educational environments is (Thai-Nghe et al. 2011): “In the educational environment, the cold-start problem is not as harmful than in the e-commerce environment where [new] users and items appear every day or even hour, thus, the models need not to be re-trained continuously.” However, this article predates the advent of MOOCs, therefore this claim is no longer true.

Among the most famous approaches to tackle the cold-start problem, one method of particular interest is an adaptive interview that presents some items to the learner, and asks the learner to rate them. Golbandi et al. (2011) build a decision tree that starts an interview process with the new user in order to quickly identify users similar to them. The best items are the ones that bisect the population into roughly two halves, and are in a way similar to discriminative items in item response theory. If we transfer this problem to adaptive assessment with test-size reduction, it can be phrased as follows: what questions should we ask to a new learner in order to infer

their whole vector of answers? The core difference with an e-commerce environment is that learners might try to game the system more than in a commercial environment, thus their answers might not fit their ability estimate.

Most collaborative filtering techniques assume that the user-to-item matrix M is of low rank r, and look for a low-rank approximation under the matrix factorization $M \simeq UV^T$ where U and V are assumed of width r. Note that, if M is binary and the loss function for the approximation is the logistic loss, we get back to the MIRT model (as a generalized linear model) described in Section 4.3.1.

Diversity Recommender systems have been criticized because they "put the user in a filter bubble" and harm serendipity. But since then, there has been more research into diversity (i.e., finding a set of diverse items to recommend), and into explained recommendations. More recently, there has been a need for more interactive recommender systems, giving more power to users by allowing them to steer the recommendations towards other directions. The application to learner systems is straightforward: this could help the learner navigate the course.

Implicit feedback In e-commerce use cases, recommender systems differentiate explicit feedback given willingly by the user, such as "this user liked this item", from implicit feedback resulting from unintentional behavior, such as "this user spent a lot of time on this page", which may imply that they are interested by the contents of this page. Such implicit feedback data is therefore used by e-commerce websites in order to know their clients better. In technology-enhanced learning use cases, explicit feedback data is often sparse; thus, implicit feedback techniques are attractive candidates to improve recommender performance. For instance, these techniques could use the time spent on a page, the search terms provided by the user, information about downloaded resources, and comments posted by the user (Verbert et al. 2011). Such data may also be useful if they are recorded while the test is administered, e.g., some course content a learner is browsing while attempting a low-stakes adaptive assessment might be useful for other learners.

Adding external information Some recommender systems embed additional information in their learning models: for instance, the description of the item, or even the musical content itself in the scope of music recommendation. In order to improve prediction over the test, one could consider extracting additional features from the problem statements of the items, and incorporate them within the feature vector.

4.3.5 Adaptive strategies for exploration–exploitation tradeoff

In some applications, one wants to maximize a certain objective function while asking questions. This leads to an exploration–exploitation trade-off: we can increase our knowledge of the user more, by exploring the space of items, or we can exploit what we know in order to maximize a certain reward. Clement et al. (2015) applied

these techniques to intelligent tutoring systems: they personalize sequences of learning activities in order to uncover the knowledge components of the learner while maximizing the user's learning progress, as a function of the performance over the latest tasks. They use two models based on multi-armed bandits: the first one relies on Vygotsky's zone of proximal development (Vygotsky 1980) under the form of a dependency graph, the second one uses an expert-specified q-matrix. They tested both approaches on 400 real students between 7 and 8 years old. Quite surprisingly, they discovered that using the dependency graph yielded better performance than using the q-matrix. Their technique helped improve learning for populations of students with larger variety and stronger difficulties.

4.3.6 Multistage testing

So far, we always assumed that questions were asked one after another. However, the first ability estimate, using only the first answer, has high bias. Thus, ongoing psychometrics research tends to study scenarios where we ask pools of questions at each step, performing adaptation only once sufficient information has been gathered. This approach has been referred to as *multistage testing* (MST) (Yan et al. 2014). After the first stage of k_1 questions, according to their performance, the learner moves to another stage of k_2 questions that depend only on their performance, and so on, see Fig. 4.6. MST presents another advantage: the learner can revise their answers before moving to the next stage, without the need of complicated models for response revision (Han 2013, Wang et al. 2015). In the language of clinical trials, MST design can be viewed as a *group sequential* design, while a CAT can be viewed as a *fully sequential* design. The item selection is performed automatically, but all stages of questions can be reviewed before administration (Chang, 2014). Wang et al. (2016) suggest to ask a group of questions at the beginning of the test, when little information about learner ability is available, and progressively reduce the number of questions of each stage in order to increase opportunities to adapt. Also, asking questions in pools means that we can do content balancing at each stage, instead of jumping from one knowledge component to the other after every question.

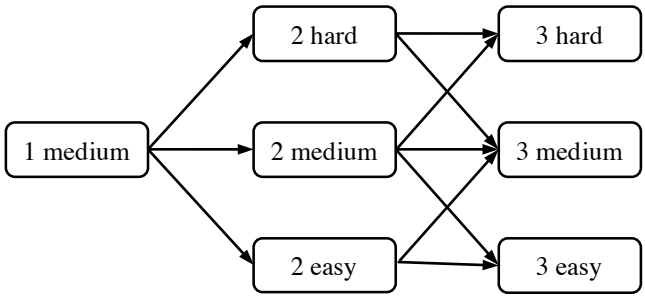

Fig. 4.6 In multistage testing, questions are asked in a group sequential design.

4.4 Comparison of adaptive testing models

Adaptive assessment models need to be validated on real data, in order to guarantee that the model accurately assesses the constructs that it is supposed to assess (Desmarais and Baker 2012). A common way to validate a model is to measure how well the assessment can predict future performance within the learning system.

To evaluate on real data the models that we presented above, we can embed them in a unified framework: all of them can be seen as decision trees (Ueno and Songmuang 2010, Yan et al. 2014), where nodes are possible states of the test, and edges are followed according to the answers provided by the learner, like a flowchart. Thus, within a node, we have access to an incomplete response pattern, and we want to use our student model and infer the behavior of the learner over the remaining questions. The best model is the one that classifies the remaining outcomes with minimal error.

Formally, let us consider a set I of students who answer questions from a set Q. Our student data is a binary matrix D of size $|I|\times|Q|$, where D_{iq} is 1 if student i answered question q correctly, 0 otherwise. An adaptive test can be formalized as follows.

TEST(student $i \in I$):
    While some questions still need to be asked
        ASK to student i the next question

We want to compare the predictive power of different adaptive testing algorithms that model the probability of student i solving question j. Thus, for our cross-validation, we need to define:

- a student training set $I_{train} \subset I$;
- a student testing set $I_{test} \subset I$;
- a question validation set $Q_{val} \subset Q$.

We use the same sets for all the models that we study. Model evaluation is performed using the EVALUATEMODEL function:

EVALUATEMODEL(model M, students I_{train}, students I_{test}, questions Q_{val}) :
    TRAIN model using lines I_{train} of D
    For each student i of I_{test} **do**
        While not all questions $\in Q \setminus Q_{val}$ have been asked
            CHOOSENEXTITEM and ask it to student i
            Evaluate predictions of model M over questions Q_{val}.

We make a cross-validation of each model over 10 subsamples of students and 4 subsamples of questions (these constant values are parameters that may be changed). Thus, if we number student subsamples I_i for $i = 1, \dots, 10$ and question subsamples Q_j for $j = 1, \dots, 4$, experiment (i, j) consists in the following steps:

- 1. train the evaluated model over all student subsamples except the i-th, i.e., $I_{train} = I \setminus I_i$;
- 2. simulate adaptive tests on the i-th student subsample (i.e., $I_{test} = I_i$) using all question subsamples except the j-th (namely, Q_j), and evaluate after each question the error of the model over the j-th question subsample (i.e., $Q_{val} = Q_j$).

The error is given by Equation (4.8), called *score* or *log loss*:

$$e(p,t) = \frac{1}{|Q_{val}|} \sum_{k \in Q_{val}} t_k \log p_k + (1 - t_k) \log (1 - p_k) \tag{4.8}$$

where p is the predicted outcome over all $|Q|$ questions and t is the true response pattern.

In order to visualize the results, errors computed during experiment (i,j) are stored in a matrix of size 10×4. Thus, computing the mean error for each column, we can see how models performed on a certain subset of questions, see Fig. 4.7.

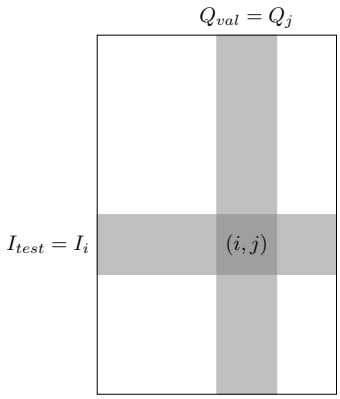

Fig. 4.7 Cross-validation over 10 student subsamples and 4 question subsamples. Each case (i,j) contains the results of the experiment (i,j) for student test set ($I_{test} = I_i$) and question validation set ($Q_{val} = Q_j$).

4.5 Results

For our experiments, we used three real datasets. The models considered were the Rasch model, the DINA model with an expert-specified q-matrix, and the GenMA model with the same q-matrix.

We now describe the results of our cross-validation, for different sizes of training and testing sets. For each dataset, the mean error of each model has been computed over all experiments.

4.5.1 ECPE

This student dataset is a 2922 × 28 binary matrix representing the results of 2922 learners over 28 English questions from the Examination for the Certificate of Proficiency in English (ECPE). The ECPE purports to measure three attributes, therefore the corresponding q-matrix has only 3 skills: knowledge of morphosyntactic rules, cohesive rules, and lexical rules. This dataset is featured in (Templin and Bradshaw 2014).

For this dataset, there were 5 student subsamples, and 4 question subsamples, i.e., the student training set was composed of 80% of the students, and the validation question sets were composed of 7 questions. The results are given in Table 4.1 and Fig. 4.8.

Table 4.1 Mean error of the different models considered for the ECPE dataset. The lowest values are denoted in bold.

Model	After 5 questions	After 10 questions	After 15 questions
Rasch	0.533 ± 0.010	0.522 ± 0.009	0.514 ± 0.009
DINA	0.535 ± 0.008	0.527 ± 0.008	0.521 ± 0.008
GenMA	**0.515 ± 0.008**	**0.484 ± 0.008**	**0.480 ± 0.008**

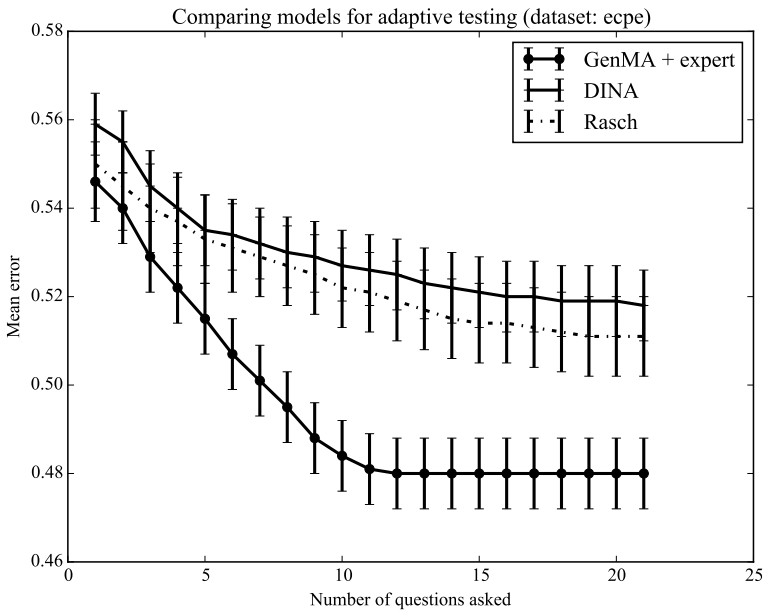

Fig. 4.8 Mean error (negative log-likelihood) over the validation question set as a function of how many questions have been asked, for the ECPE dataset.

The GenMA model outperforms the Rasch and DINA models. The estimated slip and guess parameters of the DINA model for this dataset are reported in Table 4.2.

Table 4.2 The q-matrix used for the ECPE dataset, together with the guess and slip parameters, and the success rate for each question. The highest guess value is represented in bold.

q-matrix					success rate
entries			guess	slip	
1	1	0	0.705	0.085	80%
0	1	0	0.724	0.101	83%
1	0	1	0.438	0.266	57%
0	0	1	0.480	0.162	70%
0	0	1	0.764	0.040	88%
0	0	1	0.717	0.066	85%
1	0	1	0.544	0.085	72%
0	1	0	0.802	0.040	89%
0	0	1	0.534	0.199	70%
1	0	0	0.483	0.163	65%
1	0	1	0.556	0.099	72%
1	0	1	0.195	0.305	43%
1	0	0	0.633	0.122	75%
1	0	0	0.517	0.212	65%
0	0	1	0.749	0.040	88%
1	0	1	0.549	0.126	70%
0	**1**	**1**	**0.816**	**0.058**	**88%**
0	0	1	0.729	0.086	84%
0	0	1	0.473	0.150	71%
1	0	1	0.239	0.295	46%
1	0	1	0.621	0.097	75%
0	0	1	0.322	0.188	63%
0	1	0	0.637	0.075	81%
0	1	0	0.313	0.322	53%
1	0	0	0.512	0.272	61%
0	0	1	0.555	0.211	70%
1	0	0	0.265	0.369	44%
0	0	1	0.659	0.086	81%

4.5.2 Fraction

This student dataset is a 536 × 20 binary matrix representing the results of 536 middle school students over 20 fraction subtraction questions. The corresponding q-matrix has 8 skills, described in Fig. 4.3 and can be found in (DeCarlo 2010).

There were 5 student subsamples, and 4 question subsamples, i.e., the student training set was composed of 80% of the students, and the validation question sets were composed of 5 questions. For this dataset only, we compared two occurrences of the GenMA model, one with the original expert q-matrix, the other one with a different q-matrix which was computed automatically, using sparse PCA. The results are given in Table 4.2 and Fig. 4.9.

The best model is GenMA + expert: 4 questions over 15 are enough to provide a feedback that predicts correctly 4 questions over 5 in average in the validation set.

As an example, for one of the test students, GenMA chooses 4 questions to ask in an adaptive way, then predicts that the student will correctly answer the questions from the validation question set with probabilities [61.7%, 12.3%, 41.8%, 12.7%, 12%]. Actually the true performance of the student over the validation question set is [correct, incorrect, correct, incorrect, incorrect], so the mean error is 0.350, according to equation (4.8).

Table 4.2 Mean error of the different models considered for the Fraction dataset. The lowest values are denoted in bold.

Model	After 4 questions	After 10 questions	After 15 questions
Rasch	0.461 ± 0.028	0.420 ± 0.027	0.413 ± 0.027
GenMA + expert	**0.454 ± 0.022**	**0.357 ± 0.018**	**0.322 ± 0.017**
GenMA + auto	0.544 ± 0.013	0.447 ± 0.012	0.393 ± 0.011
DINA	0.578 ± 0.030	0.429 ± 0.027	0.414 ± 0.029

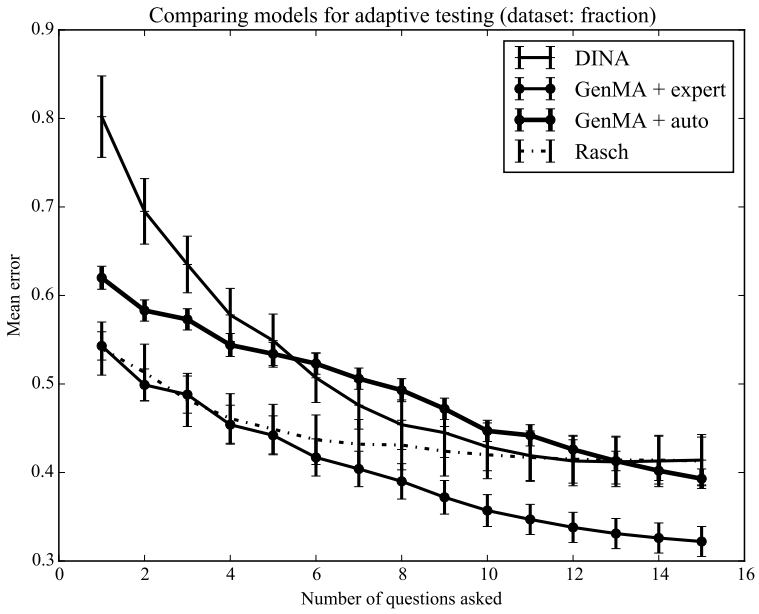

Fig. 4.9 Mean error (negative log-likelihood) over the validation question set as a function of how many questions have been asked, for the Fraction dataset.

4.5.3 TIMSS

This student dataset is a 757 × 23 binary matrix representing the results of 757 students over 23 questions from the Trends in International Mathematics and Science Study (TIMSS) 2003, an U.S. eighth grade mathematics test. The corresponding q-matrix has 13 skills over the 15 specified in (Su et al. 2013), i.e., all skills except the 10^{th} and the 12^{th}.

There were 4 student subsamples, and 2 question subsamples, i.e., the student training set was composed of 75% of the students, and the two validation question sets were composed of 11 and 12 questions. The results are given in Table 4.3 and Fig. 4.10. The best model is GenMA: after having asked 4 questions, GenMA outperforms the other models.

Table 4.3 Mean error of the different models considered for the TIMSS dataset. The lowest values are denoted in bold.

Model	After 4 questions	After 8 questions	After 11 questions
Rasch	0.576 ± 0.008	0.559 ± 0.008	0.555 ± 0.008
DINA	0.588 ± 0.005	0.570 ± 0.006	0.566 ± 0.006
GenMA	**0.537 ± 0.006**	**0.505 ± 0.006**	**0.487 ± 0.006**

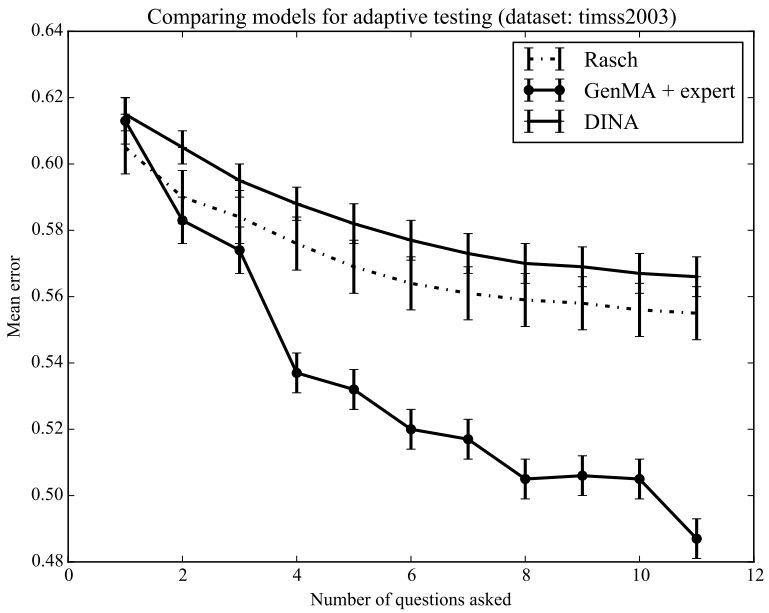

Fig. 4.10 Mean error (negative log-likelihood) over the validation question set as a function of how many questions have been asked, for the TIMSS dataset.

4.6 Discussion

In all experiments, the hybrid model GenMA with the expert q-matrix performs the best.

In the ECPE dataset, DINA and Rasch have similar predictive power, which is quite surprising given that Rasch does not require any domain knowledge. This may be because, in this dataset, there are only 3 skills: thus, the number of possible states for a learner is $2^3 = 8$, for many possible response patterns (2^{28}). Consequently, the estimated guess and slip parameters are very high (see Table 4.2), which explains why the information gained at each question is low. Indeed, the item which requires KC 2 and 3 is really easy to solve (88% success rate), even easier than items that require only KC 2 or only KC 3. Hence, the only way for the DINA model to express this behavior is to boost the guess parameter. On the contrary, GenMA calibrates one difficulty value per knowledge component, so it is a more expressive model. The same reason may explain why the mean error of GenMA converges after 11 questions: this 3-dimensional model may not be rich enough to understand the dataset, while in the Fraction dataset, the 8-dimensional GenMA model can learn after every question.

In the Fraction dataset, the DINA model tries to identify the latent state of the learner over 2^8 possible states, asking questions over few KCs at each step. This may explain why DINA requires many questions in order to converge. Rasch and GenMA-expert have similar predictive power in the early questions, but at least GenMA-expert can provide useful feedback, whereas Rasch cannot. The automatically generated q-matrix used in GenMA-auto has lower predictive power. Hence, for this dataset, Rasch provides a better adaptive assessment model than a q-matrix that is computed automatically.

In the TIMSS dataset, the DINA model tries to identify the latent state of the learner over 2^{13} possible states, which is why it needs many questions in order to reduce the prediction error. Similarly, the unidimensional Rasch model might not be enough to comprehend this multidimensional dataset. The hybrid model GenMA outperforms the other models, and can provide a feedback over 13 dimensions that achieves a mean accuracy of 77% over the validation question set of 12 questions, after 4 questions have been asked in an adaptive way.

4.6.1 Adaptive pretest at the beginning of a course

At the beginning of a course, we have to fully explore the knowledge of the learner, in order to identify their latent knowledge using as few questions as possible. This is a cold-start problem, where we have to identify whether the learner holds the prerequisites of the course, and possibly their weak and strong points. If a dependency graph is available, we suggest to use Doignon and Falmagne's adaptive assessment model (see Section 4.3.3). If a q-matrix is available, we suggest to use the

GenMA model (see Section 4.3.2). Otherwise, the Rasch model at least provides a way to measure the level of the learner in order to detect students that will require more attention.

4.6.2 Adaptive test at the middle of a course

Learners often wish to have a taste of the tasks they will be expected to solve in the final test, in the form of a mock self-assessment that does not count towards their final results. There are several scenarios to consider. If learners have access to the course while taking this low-stakes test, an adaptive assessment should take into account the fact that the level of learners may change while they are taking the test, for example because they are checking the course material during the test. Hence, this is a good use case for models that measure the progress of the learner, such as multi-armed bandits (Clement et al. 2015), mentioned in Section 4.3.5. Recall that such models need either a dependency graph or a q-matrix. If learners do not check the course material while taking the test, for example because they have limited time, the GenMA model can ask them a few questions and provide feedback, under the condition that a q-matrix is available.

Depending on the context, students should be tracked from one occurrence of the test to the next one, or not. If the test is fully anonymous, a student might get the same item twice when taking the test twice. Also they will have to record their progress themselves, e.g., by exporting their results. If students are tracked, the teacher can be notified whenever a student struggles at obtaining some KC.

Whenever students want to practice specific KCs, they can filter at the beginning of the test the KCs for which they will be assessed. This is an example of the adaptability of such models instead of pure adaptivity, as stated in Section 4.2. Students can therefore learn at their own pace.

4.6.3 Adaptive test at the end of a course

A high-stake test at the end of the course might rely on the usual adaptive assessment strategies in item response theory, in order to measure examinees effectively and grade them. On this last examination, we assume that feedback is not so useful, so any model will be suitable. Examples include the GMAT and GRE standardized tests.

4.6.4 Other applications

Adaptive assessment cognitive models such as DINA or GenMA can provide feedback under the form of degrees of proficiency over several KCs. Whenever a learner wants to sit for an anonymous test, he can understand what he did wrong. Combined with a recommender system, our model could automatically suggest lessons based on the KCs that need further work. A teacher can map student learning outcomes to KCs and KCs to items in order to be notified whenever a student is experiencing difficulty at attaining a concept. All the data collected by tests can be embedded in dashboards for visualization, in order to figure out what KCs are the most difficult to obtain for a population of students, and possibly suggest grouping students with similar difficulties, or at the contrary with disjoint difficulties.

4.7 Limitations

Here we only considered assessment of knowledge and no other dimension such as perseverance, organization, carefulness, responsibility. By reducing items, we reduce the time spent by students being assessed, which prevents boredom and leaves more time for other activities.

In our case, within a test our models never ask the same question twice. In many scenarios though, presenting the same item several times is better, for example in vocabulary learning. Spaced repetition systems based on flashcards such as Anki have been successfully used for vocabulary learning (Altiner 2011). In our case, we prefer to ask different items that need similar KCs (knowledge components), e.g., variants of a same exercise in mathematics. Such an approach has been referred to as interleaved practice (Dunlosky et al. 2013) and reduces the risks of guessing the correct answer.

Our approach is mainly static, which means we assume that the knowledge of the student does not increase within a test, even while he gets several opportunities of being assessed on the same KCs. This assumption can be made because the learner receives feedback only at the end of the test. Thus, our diagnostic test provides a snapshot of the student's knowledge at a certain time. Students can record these snapshots in order to visualize their own progress.

For simplicity, we do not consider learner metadata in our experiments, such as demographic information. This allows us to provide an anonymous test, i.e., the results are stored anonymously. This prevents stress from the examinee and helps them jump more easily into practice testing, which is useful for their learning (Dunlosky et al. 2013).

4.8 The future of assessment

We presented several models that could be used for adaptive assessment. A promising application is low-stakes adaptive formative assessments: before high-stakes assessments, learners like to train and to measure what they must know to complete the course. Such adaptive tests would be able to quickly identify the components that need further work and help the learner prepare for the final high-stakes test. It would be interesting to combine this work with automatic item generation. Learners could obtain as many variants of the same problem as they need so as to master the skills involved. The results of these adaptive tests may be recorded anonymously, so that the student can start over "with a clean state", without any tracking. Indeed, no learner would like their mistakes to be recorded for their entire lives (Executive Office of the President 2014).

With the help of learning analytics, *explicit testing* may be progressively replaced with *embedded assessment*, using multiple sources of data to predict student performance and tailor education accordingly (Shute et al. 2016, Redecker and Johanessen 2013). Indeed, if the learner is continuously monitored by the platform and if a digital tutor can answer their questions and recommend activities, they can be full actors of their continually changing progress and there is no need for an explicit test at the end of the course.

Even in such cases, however, we will still need adaptive pretests for specific uses, e.g., for international certifications (GMAT, GRE), or for newcomers at the beginning of a course, in order to identify effectively the latent knowledge they acquired in their past experience (Baker and Inventado 2014, Lynch and Howlin 2014).

Note that the only input to our adaptivity rules are the answers given so far by the learner, not their previous performance: this allows a learner to start from scratch whenever they wish. Using profile information such as the country to select the questions may lead to more accurate performance predictions: for example, from one country to another, the way to compute divisions is not the same. However, if we bias the assessment by sensitive information of this kind, we may inadvertently discriminate against some students.

In the future, an online platform could first ask the learner about their presumed knowledge. The platform could then verify if the self-assessment holds, and, if needed, explain the discrepancy. The learner could then possibly correct this assessment by proving that they actually master the knowledge components required: this could also allow them to learn more material.

4.9 Conclusion

We presented several recent student models that can be used to leverage former assessment data in order to provide shorter, adaptive assessments. As Rupp et al.

(2012), rather than attempting to determine the best model for all uses, we have compared them in terms of brevity and predictive power, to see which model is better suited to which use. Note that, throughout this chapter, we have focused on the assessment on a single learner. Readers interested in computer-supported collaborative learning in group assessments may consider reading (Goggins et al. 2015).

Models which use q-matrices are usually validated using simulated data. In this chapter, we compared the strategies on real data. Our experimental protocol could be tried on yet other adaptive assessment models. It could also be generalized to evaluate multistage testing strategies.

According to the purpose of the test (e.g., beginning, middle or end of term), the most suitable model is not the same. In order to choose the best model, one should wonder: What knowledge do we have about the domain (dependency graph, q-matrix)? Is the knowledge of the learner evolving while they are taking the test? Do we want to estimate the knowledge components of the learner or do we want to measure their learning progress while they are taking the test?

The models we described in this chapter have been introduced in several lines of work which are mostly independent. In our opinion, this implies that experts should communicate more across fields, in order to avoid giving different names to the same model. There is a need for more interdisciplinary research, and methods from learning analytics and CAT should be combined in order to get richer and more complex models. Also, crowdsourcing techniques could be applied in order to harvest more data. One might imagine the following application of implicit feedback: “In order to solve this question, you seem to have spent a lot of time over the following lessons: [the corresponding list]. Which ones helped you answer this question?” Such data can help other learners who may experience difficulties over the same questions in the future.

As we stated in Section 4.2, we think more research should be done in interactive learning analytics models, giving more control back to the learner. In this chapter, we took a first step in this direction, being inspired by CAT strategies.

The focus on modern learning analytics for personalization does not only lead to automated adaptation: it can also increase the engagement and affect of learners in the system. This raises an open question on whether the platform should let users access everything it knows about them. One advantage would be to leverage trust and engagement, one risk would be that learners may change their behavior accordingly, to try to game the system.

There exist different interfaces for assessment such as serious games or stealth assessment, which lead to more motivation and engagement from the students, e.g., Packet Tracer for learning network routing (Rupp et al. 2012), or Newton's Playground for learning physics (Shute et al. 2013). We believe our approach is more generic: it only needs student data under the form of 1 and 0's and may also be applied to these serious-game scenarios. We leave this for further research.

Acknowledgments We thank Antoine Amarilli and Ryan Lahfa for their most valuable comments. 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

Altiner C (2011) Integrating a computer-based flashcard program into academic vocabulary learning. Master's thesis, Iowa State University

Baker RS, Inventado PS (2014) Educational data mining and learning analytics. In: Learning Analytics, Springer, pp 61–75

Barnes T (2005) The q-matrix method: Mining student response data for knowledge. In: American Association for Artificial Intelligence 2005 Educational Data Mining Workshop

Bergner Y, Droschler S, Kortemeyer G, Rayyan S, Seaton D, Pritchard DE (2012) Model-based collaborative filtering analysis of student response data: Machinelearning item response theory. International Educational Data Mining Society

Chang HH (2014) Psychometrics behind computerized adaptive testing. Psychometrika pp 1–20

Chatti MA, Dyckhoff AL, Schroeder U, Thüs H (2012) A reference model for learning analytics. International Journal of Technology Enhanced Learning 4(5-6):318–331

Cheng Y (2009) When cognitive diagnosis meets computerized adaptive testing: CD-CAT. Psychometrika 74(4):619–632

Chen S, Choi A, Darwiche A (2015) Computer adaptive testing using the same-decision probability. In: 12th Annual Bayesian Modeling Applications Workshop (BMAW)

Cheng Y (2009) When cognitive diagnosis meets computerized adaptive testing: CD-CAT. Psychometrika 74(4):619–632

Clement B, Roy D, Oudeyer PY, Lopes M (2015) Multi-armed bandits for intelligent tutoring systems. JEDM-Journal of Educational Data Mining 7(2):20–48

Graduate Management Admission Council (2013) Profile of GMAT Candidates – Executive Summary. Tech. rep., Graduate Management Admission Council. http://www.gmac.com/market-intelligence-and-research/research-library/gmat- test-taker-data/profile-documents/2013-profile-of-gmat-candidates-executive- summary.aspx. Accessed 1 Apr 2016

Davier M (2005) A general diagnostic model applied to language testing data. ETS Research Report Series 2005(2):i–35

DeCarlo LT (2010) On the analysis of fraction subtraction data: The dina model, classification, latent class sizes, and the q-matrix. Applied Psychological Measurement

Desmarais MC, Baker RS (2012) A review of recent advances in learner and skill modeling in intelligent learning environments. User Modeling and User-Adapted Interaction 22(1-2):9–38

Desmarais MC, et al (2011) Conditions for effectively deriving a q-matrix from data with non-negative matrix factorization. In: 4th International Conference on Educational Data Mining, EDM, pp 41–50

Doignon JP, Falmagne JC (2012) Knowledge spaces. Springer Science & Business Media

Dunlosky J, Rawson KA, Marsh EJ, Nathan MJ, Willingham DT (2013) Improving students learning with effective learning techniques promising directions from cognitive and educational psychology. Psychological Science in the Public Interest 14(1):4–58

Executive Office of the President, Podesta J (2014) Big data: seizing opportunities, preserving values. Tech. rep., The White House

Falmagne JC, Cosyn E, Doignon JP, Thiéry N (2006) The assessment of knowledge, in theory and in practice. In: Formal concept analysis, Springer, pp 61–79

Ferguson R (2012) Learning analytics: drivers, developments and challenges. International Journal of Technology Enhanced Learning 4(5-6):304–317

Goggins SP, Xing W, Chen X, Chen B, Wadholm B (2015) Learning analytics at" small" scale: Exploring a complexity-grounded model for assessment automation. J UCS 21(1):66–92

Golbandi N, Koren Y, Lempel R (2011) Adaptive bootstrapping of recommender systems using decision trees. In: Proceedings of the fourth ACM international conference on Web search and data mining, ACM, pp 595–604

Hambleton RK, Swaminathan H (1985) Item response theory: Principles and applications, vol 7. Springer Science & Business Media

Han KT (2013) Item pocket method to allow response review and change in computerized adaptive testing. Applied Psychological Measurement 37(4):259–275

Huebner A (2010) An overview of recent developments in cognitive diagnostic computer adaptive assessments. Practical Assessment, Research & Evaluation 15(3):7

Kickmeier-Rust MD, Albert D (2015) Competence-based knowledge space theory. Measuring and Visualizing Learning in the Information-Rich Classroom p 109

Koedinger KR, McLaughlin EA, Stamper JC (2012) Automated student model improvement. International Educational Data Mining Society

Lan AS, Waters AE, Studer C, Baraniuk RG (2014) Sparse factor analysis for learning and content analytics. The Journal of Machine Learning Research 15(1):1959–2008

Leighton JP, Gierl MJ, Hunka SM (2004) The attribute hierarchy method for cognitive assessment: A variation on tatsuoka's rule-space approach. Journal of Educational Measurement 41(3):205–237

van der Linden WJ, Glas CA (2010) Elements of adaptive testing. Springer

Lynch D, Howlin CP (2014) Real world usage of an adaptive testing algorithm to uncover latent knowledge

Mandin S, Guin N (2014) Basing learner modelling on an ontology of knowledge and skills. In: Advanced Learning Technologies (ICALT), 2014 IEEE 14th International Conference on, IEEE, pp 321–323

Manouselis N, Drachsler H, Vuorikari R, Hummel H, Koper R (2011) Recommender systems in technology enhanced learning. In: Recommender systems handbook, Springer, pp 387–415

Mislevy RJ, Behrens JT, Dicerbo KE, Levy R (2012) Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. JEDM-Journal of Educational Data Mining 4(1):11–48

Peña-Ayala A (2014) Educational data mining: A survey and a data mining-based analysis of recent works. Expert systems with applications 41(4):1432–1462 Reckase M (2009) Multidimensional item response theory, vol 150. Springer

Reckase M (2009) Multidimensional item response theory, vol 150. Springer

Redecker C, Johannessen Ø (2013) Changing assessment towards a new assessment paradigm using ICT. European Journal of Education 48(1):79–96

Rupp A, Levy R, Dicerbo KE, Sweet SJ, Crawford AV, Calico T, Benson M, Fay D, Kunze KL, Mislevy RJ, et al (2012) 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(1):49–110

Shute VJ, Ventura M, Kim YJ (2013) Assessment and learning of qualitative physics in Newton's playground. The Journal of Educational Research 106(6):423–430 Shute VJ, Leighton JP, Jang EE, Chu MW (2016) Advances in the science of assessment. Educational Assessment 21(1):34–59

Shute VJ, Leighton JP, Jang EE, Chu MW (2016) Advances in the science of assessment. Educational Assessment 21(1):34–59

Su YL, Choi K, Lee W, Choi T, McAninch M (2013) Hierarchical cognitive diagnostic analysis for TIMSS 2003 mathematics. Centre for Advanced Studies in Measurement and Assessment 35:1–71

Tempelaar DT, Rienties B, Giesbers B (2015) In search for the most informative data for feedback generation: Learning analytics in a data-rich context. Computers in Human Behavior 47:157–167

Templin J, Bradshaw L (2014) Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. Psychometrika 79(2):317–339, DOI 10.1007/s11336-013-9362-0, URL http://dx.doi.org/10.1007/s11336-013-9362-0

Thai-Nghe N, Drumond L, Horváth T, Schmidt-Thieme L, et al (2011) Multi-relational factorization models for predicting student performance. In: Proc. of the KDD Workshop on Knowledge Discovery in Educational Data, Citeseer

Toscher A, Jahrer M (2010) Collaborative filtering applied to educational data mining. KDD Cup 2010

Ueno M, Songmuang P (2010) Computerized adaptive testing based on decision tree. In: 2010 10th IEEE International Conference on Advanced Learning Technologies, IEEE, pp 191–193

Verbert K, Drachsler H, Manouselis N, Wolpers M, Vuorikari R, Duval E (2011) Dataset-driven research for improving recommender systems for learning. In: Proceedings of the 1st International Conference on Learning Analytics and Knowledge, ACM, pp 44–53

Verhelst ND (2012) Profile analysis: a closer look at the PISA 2000 reading data. Scandinavian Journal of Educational Research 56(3):315–332

Vie JJ, Popineau F, Bourda Y, Bruillard (2016) Adaptive testing with a general diagnostic model. In: Design for Teaching and Learning in a Networked World: 11th European Conference on Technology Enhanced Learning, EC-TEL 2016, Lyon, France, September 12-16, 2016, Proceedings, Springer, to appear

Vygotsky LS (1980) Mind in society: The development of higher psychological processes. Harvard university press

Wang S, Fellouris G, Chang HH (2015) Sequential design for computerized adaptive testing that allows for response revision. arXiv preprint arXiv:150101366

Wang S, Lin H, Chang HH, Douglas J (2016) Hybrid computerized adaptive testing: From group sequential design to fully sequential design. Journal of Educational Measurement 53(1):45–62

Winters T, Shelton C, Payne T, Mei G (2005) Topic extraction from item-level grades. In: American Association for Artificial Intelligence 2005 Workshop on Educational Datamining, Pittsburgh, PA, vol 1, p 3

Xu X, Chang H, Douglas J (2003) A simulation study to compare CAT strategies for cognitive diagnosis. In: annual meeting of the American Educational Research Association, Chicago

Yan D, von Davier AA, Lewis C (2014) Computerized Multistage Testing. CRC Press

Zernike K (2015) Obama administration calls for limits on testing in schools. The New York Times. http://www.nytimes.com/2015/10/25/us/obama-administration-calls-for-limits-on-testing-in-schools.html. Accessed 2 Apr 2016

Zou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. Journal of computational and graphical statistics 15(2):265–286