On-The-Fly Parameter Estimation Based on Item Response Theory in Item-based Adaptive Learning Systems

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

Shengyu Jiang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Dr. Chun Wang, Advisor
Dr. Niels G. Waller, Co-Advisor

November, 2020
Abstract

An online learning system has the capacity to offer customized content that caters to individual learner’s need and has seen growing interest from industry and academia alike in recent years. Noting the similarity between online learning and the more established adaptive testing procedures, research has focused on applying the techniques of adaptive testing to the learning environment. Yet due to the inherent difference between learning and testing, there exist some major challenges that hinder the development of adaptive learning systems. To tackle these challenges, a new online learning system is proposed which features a Bayesian algorithm that computes item and person parameters on the fly. The new algorithm is validated in two separate simulation studies and the results show that the system, while being cost-effective to build and easy to implement, can also achieve adequate adaptivity and measurement precision for the individual learner.
# Table of Contents

List of Tables ............................................................................................................... iii
List of Figures .............................................................................................................. iv

Chapter 1: Introduction .............................................................................................. 1

Chapter 2: Introduction to CAT Based on IRT ...................................................... 3
  Item Response Models .............................................................................................. 3
  Computerized Adaptive Testing ............................................................................ 9

Chapter 3: Applying CAT and IRT to Learning Systems .................................... 12
  Item Bank Construction ......................................................................................... 13
  Ability tracking ..................................................................................................... 30
  Item selection ....................................................................................................... 35

Chapter 4: Current Research on Adaptive Learning Systems ......................... 39

Chapter 5: Introduction to the New Bayesian Update Rules .............................. 46
  Theoretical Background ....................................................................................... 46
  Practical Implementation ...................................................................................... 53

Chapter 6: Simulation studies ................................................................................. 61
  Study 1 .................................................................................................................... 61
  Study 2 .................................................................................................................... 68

Chapter 7: Summary and Discussion ..................................................................... 99
  Limitations ............................................................................................................. 104
  Future Directions .................................................................................................. 108

References ................................................................................................................. 110
List of Tables

Table 6.1: Parameter recovery of linear test under different test lengths and sample sizes ........................................................................................................................................................................ 65

Table 6.2a: Parameter recovery of linear test when responses were generated by 3-parameter model........................................................................................................................................................................ 66

Table 6.2b: Parameter recovery of linear test when the slope parameter follows a uniform distribution........................................................................................................................................................................ 67

Table 6.2c: Parameter recovery of linear test when ability parameter follows a skewed-normal distribution........................................................................................................................................................................ 67

Table 6.3: Item exposure statistics of the random selection condition........................................ 78

Table 6.4: Item exposure statistics of the matched difficulty condition........................................ 79

Table 6.5: Item exposure statistics of the MPWI condition........................................................ 80

Table 6.6: Parameter recovery of new items for the 20-item session condition......................... 82

Table 6.7: Parameter recovery of new items for the 30-item session condition......................... 83
List of Figures

Figure 5.1: By-person design, where a user starts the session after the previous user has finished................................................................................................................................. 54

Figure 5.2: By-item design, where an item is presented after every learner has finished the previous item................................................................................................................................. 55

Figure 5.3 Flowchart of Parameter Update Using the Bayesian Algorithm .................. 60

Figure 6.1a: Bias of ability estimates over 4 sessions in the 20-item session condition .. 86

Figure 6.1b: Bias of ability estimates over 4 sessions in the 30-item session condition .. 87

Figure 6.2a: RMSE of ability estimates over the 4 sessions in the 20-item session condition ................................................................................................................................. 89

Figure 6.2b. RMSE of ability estimates over the 4 sessions in the 30-item session condition ................................................................................................................................. 90

Figure 6.3a: Correlation between true and estimated abilities over 4 sessions in the 20-item session condition ................................................................................................................................. 92

Figure 6.3b: Correlation between true and estimated abilities over 4 sessions in the 30-item session condition ................................................................................................................................. 93

Figure 6.4a: Trace plot of the estimates and the standard deviation of a randomly chosen student’s ability under the MPWI in 20-item sessions with 50% pre-calibrated items condition ................................................................................................................................. 96

Figure 6.4b: Trace plot of the estimates and the standard deviation of a randomly chosen student’s ability under the MPWI in 30-item sessions with 50% pre-calibrated items condition ................................................................................................................................. 97
Chapter 1: Introduction

Modern technologies have made computer-assisted learning increasingly popular. Incorporating a computerized component into the learning system allows for innovative procedures to facilitate the learning process. One of the major benefits of these online learning systems is adaptivity. Compared to traditional classroom instruction where every learner receives the same materials, performs the same tasks, and solves the same problems, a computerized adaptive learning system is able to provide customized course content, format and practice sessions to match the need of individual learners (Wauters, Desmet, & Van Den Noortgate, 2010). For instance, in an online course on programming, a learner with previous coding experience would be directed to the in-depth learning modules while learners without such background would start from the more basic ones. This would prevent the advanced learners from wasting time on materials they are already familiar with and keep them motivated throughout the entire course.

The same idea can also be applied to the practice and problem-solving component of the online learning systems. Recently, there has been an increasing amount of research (e.g., Conejo, Guzmán, & Trella, 2016; Klinkenberg, Straatemeier, & Van Der Maas, 2011) dedicated to the development of such systems that deliver personalized items (sometimes with hint and/or feedback) to learners in the hope that students would gain knowledge and improve their ability through interacting with these items. To this end, the system is typically expected to be able to judge the learners’ ability level (and sometimes the difficulty of practice items), and adaptively present items that can optimize the individual learning process.
Although the development of these online practice systems is still in an early phase, there has been some research effort attempting to transfer the expertise from educational testing to the learning environment. Computerized adaptive testing (CAT) is a well-established procedure widely used in the testing industry and is very similar to the adaptive learning systems in many ways. However, the differences between learning and testing environments make some of the techniques of CAT not readily applicable to the learning scenarios. The research presented in this dissertation attempts to make a connection between adaptive testing and learning assessment by proposing an online adaptive learning system that borrows heavily from CAT designs yet employs a new Bayesian estimation algorithm that caters to the specifics of the learning environment. The new algorithm is an easy-to-implement method with closed-form equations that can compute fast and reasonable parameter estimates on the fly, which is ideal for low-stakes assessment applications like online learning.

The structure of the dissertation is outlined as follows. Chapter 2 presents a brief overview of CAT based on the item response theory (IRT) model, followed by Chapter 3 which discusses the major issues and challenges of implementing techniques of IRT and CAT in learning systems. The next chapter reviews the literature to date on the research and applications of these learning systems. Chapter 5 provides a theoretical account of the new estimation algorithm and how to implement it in the adaptive learning system that improves over the existing solutions. This is demonstrated by the results of two separate simulation studies described in Chapter 6. The final chapter summarizes the main findings as well as limitations in the simulations and lays out possible directions for future research.
Chapter 2: Introduction to CAT Based on IRT

Item Response Models

The unidimensional model. IRT models the interaction between an item and a person by calculating the probability of possible responses to an item from persons with a given latent trait level. The mathematical formulas of IRT usually take either one of the two forms: the logistic function or the normal ogive function. Models for dichotomously scored items can be further classified by the number of item parameters in the model. For instance, the 3-parameter logistic (3PL) model can be written as

\[ p_j(u_{ij} = 1|\theta_i) = c_j + (1 - c_j) \frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}} \]  

(1)

where \( a_j \) and \( b_j \) denote discrimination and difficulty parameters of item \( j \), respectively. \( c_j \) is the pseudo-guessing parameter that defined the lower asymptote of the item characteristic curve. \( \theta_i \) denotes the latent trait of person \( i \), and \( u_{ij} \) is the observed binary response of person \( i \) to item \( j \). The model can be reduced to the 2-parameter version by setting \( c_j \) to 0: \( p_j(u_{ij} = 1|\theta_i) = \frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}} \).

For the normal ogive model, let \( \Phi \) be the cumulative distribution function of the standard normal distribution, then the 2-parameter model is simply

\[ p_j(u_{ij} = 1|\theta_i) = \Phi\left(a_j(\theta_i - b_j)\right) \]  

(2)

The above equation can be written in slope-intercept format. Let \( \beta_j = a_j \) and \( \alpha_j = -a_j b_j \)
\[ p_j(u_{ij} = 1|\theta_i) = \Phi(\beta_j \theta_i + \alpha_j) \]  

(3)

One of the advantages of IRT over classical test theory is that it allows for different levels of measurement precision at different \( \theta \) levels. The measurement precision as a function of \( \theta \) can be quantified as observed or expected (or Fisher) information. The observed information is defined as the negative second derivatives of the log likelihood function for item \( j \) evaluated at the \( \theta \) level of the person

\[ J_{Uj}(\theta) = -\frac{\partial^2 \ell}{\partial \theta^2} \]  

(4)

where \( \ell \) is the log-likelihood function for item \( j \) evaluated at \( \theta \) (the subscript \( i \) is omitted for simplicity). The Fisher information \( I_{Uj}(\theta) \) is the expectation of \( J_{Uj}(\theta) \). For example, for 2PL binary models, the two types of information are identical, and the Fisher information is

\[ I_{Uj}(\theta) \equiv E \left( J_{Uj}(\theta) \right) = \alpha_j^2 P_j(\theta) Q_j(\theta) \]  

(5)

where \( P_j(\theta) = P(u_j = 1|\theta) \) and \( Q_j(\theta) = 1 - P_j(\theta) \).

For the normal ogive model, the observed and Fisher information are not the same as the observed information also depends on the response data. Using the slope-intercept parametrization, the Fisher information is calculated as

\[ I(\theta) = \beta_j^2 \phi(\beta_j \theta + \alpha_j)^2 \]  

\[ \frac{P_j(\theta) Q_j(\theta)}{} \]  

(6)

where \( \phi \) is the probability density function of the standard normal distribution (Baker & Kim, 2004).
The models introduced above assume binary responses (correct/incorrect) and can be extended to items with more than two response categories. For an item with \( C \) response options, the normal ogive form of the graded response model (GRM) can be written as

\[
P(u_{ij} = c | \alpha_j, \beta_j, \theta_i) = \Phi(\beta_j \theta_i + \alpha_j - \gamma_c) - \Phi(\beta_j \theta_i + \alpha_j - \gamma_{c+1})
\]  

(7)

with \( \gamma_0 = -\infty \) and \( \gamma_C = \infty \). \( \gamma_c \) is the threshold parameter when the observed response \( u_{ij} = c \).

Note that GRM can be reduced to binary model by letting \( C = 2 \) and \( \gamma_1 = 0 \). Then we have \( P(u_{ij} = 1 | \alpha_j, \beta_j, \theta_i) = \Phi(\beta_j \theta_i + \alpha_j) \) when \( c = 1 \) (correct answer) and \( P(u_{ij} = 0 | \alpha_j, \beta_j, \theta_i) = 1 - \Phi(\beta_j \theta_i + \alpha_j) \) when \( c = 0 \) (incorrect answer).

**Parameter estimation.** The item and ability parameters in IRT are usually estimated separately and sequentially with different statistical methods. In the typical scenario where item and ability parameters are both unknown, the common practice is to first estimate the item parameters and then calculate ability parameters using the item parameter estimates obtained in the previous step. To compute item parameters, the most popular method is marginal maximum likelihood estimation based on expectation-maximization (MMLE/EM) algorithm (Bock, 1972; Bock & Aitkin, 1981). Specifically, the algorithm iterates between an expectation (E) step that gives the expectation of the marginalized likelihood and a maximization (M) step that, as the name suggests, maximizes it. To find the expectations of item log-likelihood, a crucial intermediate step is to find the posterior ability distribution of each person. First, write the likelihood function of the response vector \( U_i \) from a given person as
\[ L_i(\theta) = P_i(U_i|\theta) = \prod_{j=1}^{P} P_i(\theta_j)^{u_{ij}/Q_j(\theta_j)^{1-u_{ij}}} \]  

which is conditional on the ability level (the likelihood is conditional on item parameters as well, which are omitted for simplicity). The marginal likelihood is then found by integrating out \( \theta \)

\[ L_i(U_i) = P_i(U_i) = \int P_i(U_i|\theta) \varphi(\theta) \, d\theta \]  

where \( \varphi(\theta) \) denotes the latent distribution of ability, which is often assumed to be standard normal. Applying Bayes’ theorem, the posterior distribution of \( \theta \) is

\[ P_i(\theta|U_i) = \frac{P_i(U_i|\theta)\varphi(\theta)}{P_i(U_i)} \]  

This posterior is usually evaluated using the Gaussian-Hermite quadrature rules, with \( \theta_q \) and \( A(\theta_q) \) denoting the \( q \)th quadrature point on the \( \theta \) scale and the corresponding weight, respectively. The quadrature form of Equation 10 can be written as

\[ P_i(\theta_q|U_i) = \frac{L_i(\theta_q)A(\theta_q)}{\sum_{q=1}^{Q} L_i(\theta_q)A(\theta_q)} \]  

Equation 11 is useful in computing the following two quantities: \( \bar{N}_{jq} \), the expected number of persons at the \( q \)th quadrature point that answer a given item, and \( \bar{r}_{jq} \), the expected number of persons at the \( q \)th quadrature point that answer the item correctly.

Suppose the item in question is item \( j \) and there are in total \( N_j \) persons who respond to this item. The two quantities correspond to the sum and weighted (by response) sum of the posteriors over all the persons that answer item \( j \)
\[
\tilde{N}_{jq} = \sum_{l=1}^{N_j} \frac{L_i(\theta_q) A(\theta_q)}{\sum_{q=1}^{Q} L_i(\theta_q) A(\theta_q)}
\] (12)

\[
\tilde{r}_{jq} = \sum_{l=1}^{N_j} \frac{u_{ij} L_i(\theta_q) A(\theta_q)}{\sum_{q=1}^{Q} L_i(\theta_q) A(\theta_q)}
\] (13)

The expected log-likelihood of item \( j \) can then be written as

\[
\tilde{\ell}(a_j, b_j) = \sum_{q=1}^{Q} \tilde{r}_{jq} \log P_j(\theta_q) + \left( \tilde{N}_q - \tilde{r}_{jq} \right) \log Q_j(\theta_q)
\] (14)

where \( P_j(\theta_q) \) and \( Q_j(\theta_q) \) are defined the same way as in Equation 5. In the M-step, assuming the model is in logistic form, the score functions are

\[
\frac{\partial \tilde{\ell}}{\partial a_j} = \sum_{q=1}^{Q} \left( \tilde{r}_{jq} - \tilde{N}_q P_j(\theta_q) \right) (\theta_q - b_j)
\] (15)

\[
\frac{\partial \tilde{\ell}}{\partial b_j} = \sum_{q=1}^{Q} \left( \tilde{r}_{jq} - \tilde{N}_{jq} P_j(\theta_q) \right) a_j
\] (16)

Set the score functions equal to zero and solve for the item parameters using optimization procedures such as Newton-Raphson. In some situations, such as those involving multiple group data, it is also possible to estimate the moments of the latent distribution after the item parameter estimates are obtained. The idea is similar except that now the marginal likelihood of the entire response matrix is maximized with respect to the population mean and standard deviation (SD).

\[
L(U) = \prod_{i=1}^{L} L_i(U_i)
\] (17)
Instead of employing an iterative procedure like Newton-Raphson, setting the score functions of Equation 17 to zero results in closed-form equations from which the estimates can be computed directly (Baker & Kim, 2004). For instance, the estimate of the population mean is

\[
\hat{\mu} = \frac{1}{N} \sum_{q=1}^{Q} \theta_q \sum_{i=1}^{N} P_i(\theta_q | U_i)
\]  

(18)

where N is the total sample size and \( P_i(\theta_q | U_i) \) can be found using Equation 11.

The resulting new population parameter estimates can be used to find the new quadrature points and weights in the next iteration, while the new item parameter estimates are plugged into the likelihood functions Equation 8 and Equation 14. The process is repeated until the change in estimates between two consecutive cycles is smaller than a predefined threshold, i.e. the algorithm reaches convergence.

After item parameters are known, computing ability estimates becomes straightforward. The simplest approach, maximum likelihood estimation (MLE), finds the ability estimate that maximizes the likelihood in Equation 8. In practice, the log-likelihood is used instead since it is easier to work with, and the optimization could be done with the Fisher scoring or Newton-Raphson algorithm, among others.

One issue with MLE is that it will not work given a perfect response vector, that is, the answers are either all correct or incorrect. In this case, Bayesian estimation methods should be considered. An example is expected a posteriori (EAP), which specifies a prior distribution for the ability parameter and uses the expectation of the
posterior ability distribution as the estimate. The posterior distribution is given by Equation 10 and the estimate is

$$\hat{\theta}_{EAP} = \int \frac{\theta P_i(U_i|\theta)\phi(\theta)}{P_i(U_i)} \, d\theta$$

which is evaluated using quadrature methods similar to those in Equation 12 and 13.

**Computerized Adaptive Testing**

The most important feature of CAT is to let the examinees receive items with difficulties that are tailored to their estimated ability levels. To this end, IRT is an ideal choice as the measurement model of CAT because it models item responses in terms of item and person parameters, which makes the matching of item and examinee possible. Testing programs built upon IRT and CAT have been successfully implemented in many areas and can provide more accurate and efficient measurement than traditional paper-and-pencil tests (Van der Linden & Glas, 2010; Wainer & Mislevy, 2000).

CAT procedures generally consist of the following key components:

1. Starting values of ability estimates
2. Item selection
3. Ability estimation
4. Termination criterion

When examinees first enter the CAT system, their ability levels are unknown, and starting values have to be assigned. This could either be random or based on background information gathered about the examinee, such as previous educational record, age or gender (e.g., Van Der Linden, 1999). After the ability level is initialized, an item
selection index is calculated using the initial ability and item parameters for all items available to be administered, and the candidate item best matched to the current ability level is selected and delivered. Item selection in CAT has been heavily researched and some of the most common options include matched difficulty (e.g., Ban, Hanson, Yi, & Harris, 2002), maximal Fisher information, Kullback-Leibler information (Chang & Ying, 1996), and the Bayesian item selection methods (van der Linden, 1998).

Among these, maximal Fisher information is arguably the most straightforward and widely used approach. It selects the candidate item that contains the largest amount of information calculated using Equation 6 evaluated at the current ability estimates but may not function well when the current estimate is far from the true value. A popular alternative, the Kullback-Leibler (KL) information, is a global information index that considers not just the point estimate of ability, but an interval around the current point estimate; and thus it is more robust to inaccurate estimates that often occur in early stage of CAT (Chang & Ying, 1996). The Bayesian item selection criteria differ from the above methods in that they make use of the posterior distribution of ability parameters instead of point estimates, which allows them to capture the variability of the parameters. One such criterion is the maximum posterior weighted information (MPWI, van der Linden, 1998). Essentially, it is a Bayesian extension of the maximum information criterion that weighs item information with the posterior distribution of \( \theta \):

\[
MPWI = \int J_{U_j}(\theta_i) g(\theta_i | u_{i_1}, ..., u_{i_k}) d\theta_i
\]

(20)

where \( J_{U_j}(\theta_i) \) is the observed information for candidate item \( j \) evaluated at the current ability level of person \( i \) and \( g(\theta_i | u_{i_1}, ..., u_{i_k}) \) is the posterior distribution of \( \theta_i \). The
integral can be computed using quadrature methods mentioned above or Monte Carlo integration. MPWI has been shown to perform similarly to the other more sophisticated methods proposed by van der Linden (1998) for CAT with polytomous items and was much faster to compute (Choi & Swartz, 2009; Penfield, 2006).

After the examinee responds to the selected item, it is now possible to estimate the ability level using the response vector and corresponding item parameters. This could be done using the MLE or EAP methods discussed in the previous section, or a combination of both since a perfect response pattern is quite likely in the early stage of CAT where MLE is not viable. The examinee’s ability estimate is then updated, and the same process of item selection, data collection, and ability estimation is repeated until a termination criterion is met. The test could either be finished when a fixed test length is reached, or some prespecified level of measurement precision is achieved.
Chapter 3: Applying CAT and IRT to Learning Systems

CAT and IRT have been well established in standardized testing, and it seems natural to apply the procedures to item-based adaptive practice systems, as the two systems share many common features. At the core, both CAT and the adaptive practice system deliver items adaptively according to the students’ ability levels and include similar procedures such as item selection and ability estimation. Despite the similarities, these two systems still differ in some important respects due to the different natures of testing and learning. For instance, in a learning environment it is expected that students’ ability would show some form of change (preferably improvement) over time while in testing it is assumed that ability will remain constant throughout the testing session. Another difference is that many online learning and instruction programs provide item-level feedback to the learner immediately after they respond to the item so that they can improve on their abilities as they work through the items. In testing, however, feedback is typically limited to a report of total score or subscores that is only available after the entire test is finished, and item-level feedback during the test is typically not allowed. Finally, in a learning system, the results of assessment are used to facilitate learning, not to inform high-stakes decisions with legal implications. The scores obtained from the learning assessment systems should only be considered as reasonable estimates of learner ability that should not be used in the same way as scores from large-scale standardized tests.

These differences pose technical and practical challenges when applying CAT methodology in the learning environment and have been noted in the literature (Eggen, 2012; Veldkamp, Matteucci, & Eggen, 2011; Wauters et al., 2010). The challenges can
be summarized into three broad categories: item bank calibration, dynamic ability estimation, and item selection.

**Item Bank Construction**

The first challenge involves calibrating the items that are going to be used for adaptive measurement. In CAT, an item bank consists of items with known parameters that are obtained from previous calibration studies, and the accuracy of these estimated parameters are crucial in determining the performance of subsequent steps of item selection, ability estimation, and in some instances, test termination. Research has shown that errors in item calibration could lead to issues in information-based item selection (Hambleton & Jones, 1994), as well as the bias (Doebler, 2012) and standard errors (Cheng & Yuan, 2010) of ability estimates. Therefore, the quality of the item bank is pivotal in the implementation of any CAT programs.

The process of item calibration in IRT is known to be costly and time-consuming as every item in the bank needs to be administered to a representative and motivated sample of adequate size (Wauters et al., 2010). This requires large, high-quality datasets that are oftentimes only affordable for large-scale, high-stakes testing programs (Conejo et al., 2016). Although the procedure for the item parameters used in learning applications may not need to be as rigorous as in testing (Veldkamp et al., 2011), parameter estimates are still needed in order to achieve adaptivity. Item calibration in the learning environment can be particularly challenging because more items can be included in practice and instruction than in testing. A learning program could go on for several weeks with multiple separate sessions, which would require in total many more items for a single user and as a result a larger item bank. On the other hand, in most standardized
testing programs examinees are only allowed a single sitting with a time limit in place, and the number of items administered to an individual examinee is also limited.

Researchers have adopted different approaches to address the issue of item bank calibration. The common goal is to bypass a full-scale calibration study and derive item parameter estimates from other sources. Below is a summary of the current literature dealing with this issue from different research areas.

**Heuristic approach.** Some studies attempted the use of human judgment, such as feedback from domain experts, teachers, or learners themselves, to provide estimates for item difficulty (e.g., Conejo, Guzmán, Perez-De-La-Cruz, & Barros, 2014; Wauters, Desmet, & Van Den Noortgate, 2012). Their results showed that these estimates tended to be less reliable than calibration with statistical methods and could be equally resource-demanding in the face of a large item bank. Others proposed to construct statistical models that can predict the difficulty of items based on item features. For instance, in a mobile app that offers training in English vocabulary, the difficulty of the individual words in the system was estimated from a weighted combination of word properties such as word length, number of phonemes, and their frequency of usage (Chen & Chung, 2008). Note that building these predictive models usually requires justification from substantive knowledge, which may not exist in some fields. In addition, parameter estimates obtained from this statistical analysis are only preliminary and must be validated after operational data start to accumulate.

**Automatic item generation (AIG).** Automatic item generation has been proposed as a promising alternative to the lengthy and costly process of traditional item development (Gierl & Haladyna, 2013; Gierl & Lai, 2013) At the core of AIG is the so-
called item models or parent items, which are usually operational items that are carefully selected from previous testing sessions (Embretson & Kingston, 2018; Gierl & Lai, 2013). These parent items are then broken down into several components that can be substituted with different materials. A simple example would be a question on geography asking about the capital city of France. In the item stem “France” is deemed as a variable component, and by replacing “France” with “Russia” a new item is generated. In practice both the item stem and the generative rules are much more complex, but the idea is that through manipulation of these variable components a large number of new items can be created which share similar structures to their parents. More importantly, these new items should also have similar, or at least predictable psychometric properties that allow test developers to expedite or simplify the calibration process.

Depending on whether the new items are expected to have identical psychometric properties, AIG can be divided into two subtypes (Embretson & Kingston, 2018; Geerlings, Glas, & van der Linden, 2011; Gierl & Lai, 2013). One of them is called item cloning, which refers to generating new items that are psychometrically equivalent to their parents and involve the same type and level of cognitive process to solve (Lathrop & Cheng, 2017). The other type of AIG produces structurally variant items (Embretson & Kingston, 2018) whose difficulty is varied and manipulated according to cognitive theories on problem solving and information processing (e.g., Daniel & Embretson, 2010; Embretson, 1999).

Naturally, the two types of AIG call for different statistical modeling approaches (Geerlings et al., 2011; Sinharay & Johnson, 2013). Grouping the parent item and all its clones into an “item family”, item clones are usually modeled by a multilevel approach
which assumes that all the item parameters within a family follow a joint distribution from which the parameters of item clones are random draws (Glas & van der Linden, 2003; Lathrop & Cheng, 2017; Sinharay & Johnson, 2013). The multilevel modeling approach conceptualizes the deviation of item clones from their parents as a random component to account for the dependency structure of items within the item family. Glas and van der Linden (2003) proposed a two-level model whose first level was the 3PL IRT model, and the second level specified the family distribution as multivariate normal. They also discussed the related item calibration methods and adaptive item selection rules based on their model. For the automatically generated items with varied psychometric properties, it is assumed that the differences within the item family are systematic because they are introduced by varying the underlying cognitive processes required to solve an item. The systematic variation in item parameters can be explained by treating the parameters of item clones as linear combinations of item covariates (Sinha ray & Johnson, 2013). For instance, Embretson (1999) proposed a 2PL-constrained model for the Raven’s Progressive Matrix Test and used perceptual item features, such as whether the objects were distorted or not, to predict both difficulty and discrimination parameters.

It should be noted, however, that performance of such predictive models would depend heavily on the choice of item covariates, which requires substantial input from learning specialists (Lathrop & Cheng, 2017).

AIG has been gaining popularity among test developers because it meets the growing need of producing a large number of items in a speedy, cost-efficient fashion (Gierl & Lai, 2013). The same need also exists for the online learning environments. In fact, learning may be a better scenario to apply the techniques of AIG. One of the major
criticisms of automatically generated items, especially item clones, in testing applications is that they sometimes look too similar to each other and can be easily identified if the test is administered on a continuous basis, compromising test security (Gierl & Lai, 2013). For learning, test security is less of an issue and it is actually not uncommon for students to practice similar items multiple times as a learning process.

The problem now remains as to how to obtain the item parameter when it comes to applying AIG to the learning context. For the item cloning approach, the simplest and least expensive solution would be to use the parameters of parent items for all the item clones in the family. Yet this is unlikely to hold true in practice and in fact all of the statistical models discussed earlier assumed that the generated clones would not have the exact same parameters as their parents. This means that some sort of calibration study is still needed to obtain the necessary parameter estimates. Even in the multilevel approach where finding the parameters of individual item clones is not necessary, calibration studies still need to be conducted to find the family level parameters such as the variance-covariance matrix of the joint distribution (Glas & van der Linden, 2003). That is, instead of doing away with calibration studies altogether, AIG simply calls for a calibration study of a different type, and the savings in test development brought by AIG would have more to do with item writing than item calibration.

Therefore, to streamline the procedure of item bank calibration, it is necessary to assume that automatically generated item clones have the same parameters as their parents. The impact of ignoring the deviation introduced by item cloning has been investigated in a few simulation studies (Bejar et al., 2003; Colvin, Keller, & Robin, 2016; Embretson, 1999; Lathrop & Cheng, 2017). Generally, it has been found that
ignoring the discrepancy between parent items and item clones resulted in a small yet systematic increase in the bias and/or standard error of ability estimates in the context of linear, adaptive, and multistage testing. It remains an open question whether this loss of measurement precision could justify treating all items within the same family as identical, and in practice the decision may come down to a tradeoff between cost and performance.

**Online calibration.** Another line of research on item bank calibration focuses on the idea of online calibration, a procedure in CAT that replenishes the item bank over the course of a testing program by calibrating newly added items as they are being administered in operational tests. The process for online calibration is largely the same as a regular CAT session with a group of examinees whose latent traits are usually unknown and a bank of operational items whose parameter estimates are obtained from previous calibration (Stocking, 1988). In online calibration, however, there is also a pool of new items (also called pretest items or field-test items) with unknown parameters that need to be calibrated so that they could be added to the operational bank. In order to calibrate these new items while also fulfilling the task of scoring examinees, the CAT system would select a number of new items and deliver these items at prespecified positions (called seeding locations) of a test in the exact same fashion as operational items (it is actually very important to make sure that examinees cannot tell the new items from operational items). The responses to the new items are usually not used for scoring since the parameters of new items are either unknown or unstable during testing; instead, they are only collected to calibrate or update the parameters of new items. The calibration process continues until satisfactory new item parameter estimates are obtained, and the
new items are then “promoted” to be eligible for use in the operational pool for future administrations (van der Linden & Ren, 2015; Zheng, 2016).

Research on online calibration can be broadly classified into two categories: online calibration methods and online calibration design (He, Chen, Li, & Zhang, 2017). The former addresses the statistical methods for estimating the new item parameters, while the latter focuses on design issues similar to those encountered in designing a CAT system, such as item selection and termination rules. The remainder of this section will review both categories in detail.

Estimation methods. One of the major challenges of online calibration is how to put new items on the same scale as operational items, which have been previously calibrated with a separate sample. Another major challenge is the issue of data sparseness (P. Chen, Wang, Xin, & Chang, 2017). Due to practical limits, usually each examinee can only respond to a fraction of the new items from the new item pool. For example, in one simulation study, each simulee answered 10 new items selected from a pool consisting of 240 new items, resulting in a very sparse response matrix with many missing values (J. C. Ban et al., 2002).

In one of the earliest papers on this topic, Stocking (1988) proposed two estimation methods for online calibration based on conditional maximum likelihood estimation. Her Method A used the responses to operational items to find the person parameter estimates for each individual, and then plugged in these estimates into the log-likelihood as true values to find the parameter estimates of new items. Method B attempted to correct for the error of using person parameter estimates in place of true
values by performing a scale transformation (Stocking & Lord, 1983) based on additional anchor items.

Another family of estimation method are modified versions of the MMLE/EM algorithm. The MMLE with one EM cycle (OEM) is a non-iterative algorithm that calculates in one single E-step the posterior distribution of $\theta$ based on responses to operational items only, and estimates the parameter of the new item $j$ in one single M-step based on the responses by all the $N_j$ persons who respond to that item (Wainer & Mislevy, 2000). The following marginal likelihood is maximized

$$L_j^* = \prod_{i} \int L_j(\theta_i) g(\theta_i | \alpha_{op}^i, \beta_{op}^i) d\theta_i$$  

(21)

where $L_j(\theta_i) = P_j(\theta_i)^{u_{ij}}Q_j(\theta_i)^{1-u_{ij}}$ is the likelihood of the response of person $i$ to the new item $j$, and $g(\theta_i | \alpha_{op}^i, \beta_{op}^i)$ is the posterior distribution of $\theta_i$ conditional on parameters of the operational items received by person $i$.

MMLE with multiple EM cycles (MEM) improves on the OEM method by adopting an iterative approach more similar to the original EM algorithm, allowing for multiple EM cycles until convergence is reached for the new item parameters (J.-C. Ban, Hanson, Wang, Yi, & Harris, 2001). The first EM cycle is exactly the same as OEM. Starting from the second cycle, the posterior distribution of $\theta$ is computed from the responses to operational as well as new items, whose parameter estimates are found provisionally in the first cycle. The marginal likelihood to be maximized from the second cycle on becomes
where $\hat{\alpha}_{new}^i$ and $\hat{\beta}_{new}^i$ are the vectors of estimated parameters of the new items received by person $i$.

A closely related method, BILOG with Strong Prior (J.-C. Ban et al., 2001), which is based on the computer program BILOG (Mislevy & Bock, 1990), sets the prior means of the operational items to equal to the corresponding values and a very small prior variance on the operational items since their parameters are known. It differs from MEM in that the operational items are also estimated along with the new items. In contrast, MEM only estimates the new item parameters while operational items remain untouched. After setting the strong priors for operational items and using the BILOG default priors for new items, this method proceeds with the standard MMLE/EM procedure until convergence is met.

Simulation results (Ban et al., 2001; Ban et al., 2002) showed that generally MEM was the best method overall in terms of parameter recovery, while Stocking’s Method B, although performing similarly, had the extra requirement of anchor items. BILOG with Strong Prior did not fare well in small sample sizes, and OEM was not recommended in almost every case. More recently, these methods have been successfully adapted and applied to cognitive diagnostic CAT (P. Chen, Xin, Wang, & Chang, 2012) and multidimensional CAT (P. Chen, 2017; P. Chen et al., 2017), and the findings were largely consistent with MEM performing the best.
The Bayesian version of Stocking’s Method A, OEM, and MEM were developed and compared in a simulation study, and the Bayesian version of MEM also emerged as the winner (Zheng, 2014). Another Bayesian online calibration method was proposed by van der Linden and Ren (2015) who used Markov Chain Monte-Carlo (MCMC) sampling to obtain the posterior distribution of parameters of person and new items. In their paper, the parameters of operational items were represented by vectors of random draws from the posterior distributions derived from previous calibration. These posteriors were permanently stored in the systems and resampled every time the operational item was needed for item selection or ability update. The person and new item parameters were estimated by updating their posterior distributions using the Metropolis-Hastings within Gibbs algorithm (Patz & Junker, 1999), and random draws from these updated posteriors were saved as interim estimates, overwriting the previous vectors. One issue with this method is that in the estimation process there are many tuning variables that need to be determined by pilot simulations, such as the length of the burn-in period, the variance of the proposal distribution, etc. It is questionable whether these variables are generalizable for different datasets and setups. In a follow-up study (Ren, van der Linden, & Diao, 2017), they adopted the same Bayesian framework but instead used Gauss-Hermite quadrature to evaluate the posteriors so that the fine-tuning of the MCMC procedure was no longer needed.

**Calibration Design.** There are many aspects to the online calibration design and four of them will be discussed: the parameter estimates update frequency, seeding locations, stopping rules, and selection methods of new items. Among these four factors,
new item selection has received the most research and is found to have the largest impact on calibration performance.

*Frequency of parameter update.* In online calibration, there is no consensus as to when or how often the new item parameters should be updated. Sometimes they are never updated during the test like in traditional calibration and only estimated once the entire testing session is over and all responses have been collected (Ban et al., 2002; He et al., 2017). Sometimes the test is divided into multiple stages with prespecified cutoff points, and the new items are estimated after a given stage as provisional estimates for use in the subsequent adaptive item selection (P. Chen et al., 2012; Makransky & Glas, 2010). In other simulations, the parameters are updated periodically after a batch of examinees have provided responses to the item (van der Linden & Ren, 2015; Zheng, 2016). Theoretically, it is possible, and actually ideal, to update the new item parameter as soon as a new response comes in (called a fully sequential approach), as the new item parameter would always reflect the most current information offered by the data (Ren et al., 2017; van der Linden & Ren, 2015). In practice, however, this is rarely implemented due to limiting factors such as the heavy computational burden associated with the calibration methods mentioned in the previous section. The effect of parameter update frequency was recently investigated in a simulation study by varying the batch size of new responses (Ren et al., 2017). No systematic difference was found between batch sizes of 20, 50, and 100 except for more even item exposure for the smaller batch size.

*Seeding locations.* In online calibration, seeding locations refer to the positions in a test where the new item is delivered. In online calibration, it is desirable that examinees cannot distinguish between new and operational items, and respond to both types of items
with similar level of effort and motivation (P. Chen et al., 2017; Jones & Jin, 1994). Therefore, in practice seeding locations are usually scattered randomly throughout the test rather than fixed. In addition, it has been argued that in some situations where the new items are selected adaptively (more details will be given in the new item selection section), it would potentially be preferable to present new items at a later stage of the test when ability parameters will be estimated more accurately (Ren et al., 2017; Zheng, 2016). However, results from a simulation study did not corroborate this claim and found minimal impact of seeding location on parameter recovery (Zheng, 2016).

**Stopping rules.** Analogous to CAT, stopping rules in online calibration determine when a new item has finished the calibration process and can start functioning as an operational item. In some simulation studies, there is no explicit mention of stopping rules for new items and item calibration stops once the entire testing session is completed (Ban et al., 2001; P. Chen et al., 2017; He et al., 2017). Others would continuously administer the new items to more examinees until a predetermined criterion is met (Y. I. Chang & Lu, 2010; Ren et al., 2017). One such stopping rule for calibrating new items is when a given number of persons have responded to them. Previously used choices of target sample size include 1000 for 2PL model (Ren et al., 2017) and generalized partial credit model (Zheng, 2016), and 500 for the Rasch model (Kingsbury, 2009). Another criterion is precision of the new item parameter estimates. This could be indexed by the standard error of estimates or the posterior standard deviation (Ren et al., 2017). Ren et al. (2017) used the posterior standard deviation of 0.1 for discrimination parameter and 0.06 for difficulty parameter in the 2PL model as one of their stopping criteria and compared it with the target sample size of 1000. They found that the standard deviation
rule generally required larger sample size than the threshold of 1000, yet the parameter recovery was already good enough for the latter. They suggested that the choice of stopping rule is primarily of practical concern.

*New item selection.* In online calibration, operational items are typically assigned using common item selection methods in CAT, such as maximum Fisher information (eg. Ban et al., 2001; Kingsbury, 2009; van der Linden & Ren, 2015), matched difficulty (eg. Ban et al., 2002), and Bayesian criteria like A-optimality (eg. Chen et al., 2017). The assignment of new items is more complicated because the item parameters of new items are usually unknown or contain large errors and unless they are to be delivered randomly, the item selection indices cannot be calculated. Some form of initialization is needed to obtain provisional parameter estimates for the new items. Another aspect of the issue is that, as is discussed in seeding locations, even though new items are not used in scoring, it is important that examinees respond to them as if they were operational items. Therefore, selecting new items that do not stand out from surrounding operational items would help secure higher quality data for calibration.

The simplest way of assigning new items during an operational CAT is randomly selecting new items from the pool. This method does not require initial parameter estimates of new items and could make sure that each new item is administered to an approximately equal number of examinees. This is important because an adaptive item selection criterion may underexpose some items in the pool so that these items would either not be as well calibrated as others or, in the case of using stopping rules, remain in calibration phase for too long (Makransky & Glas, 2010; Ren et al., 2017). The downside of random selection is that it would sometimes select items that significantly differ from
surrounding items in terms of difficulty, especially at the later stage of the test where items are concentrated at the person’s true ability level (P. Chen et al., 2017; Jones & Jin, 1994; Kingsbury, 2009).

To overcome this issue in random calibration, adaptive online calibration used the same item selection method for both operational and new items. Because the new item parameters are not initially available, Kingsbury (2009) suggested using individual ratings or item characteristics to estimate item statistics before any test data was collected, similar to the heuristic approaches presented in Wauters et. al. (2012). However, he only studied the Rasch model and it is unclear how these provisional estimates would work for item discrimination. Another approach divided the test into two or more stages (P. Chen et al., 2017, 2012; Makransky & Glas, 2010; Zheng, 2016). In the first stage, new items are randomly selected, and the item parameters were obtained at the end of this stage. In the following stages, new items are adaptively administered at their seeding locations in the same way as operational items with item selection indices calculated from the provisional parameter estimates derived from the first stage.

Note that because responses to new items are not used for scoring, adaptive online calibration does not enhance efficiency in estimating person ability. The benefit of adaptive online calibration is mostly limited to its ability to conceal the new items from examinees. Actually, it has been observed that in some cases random selection actually produced more accurate item parameter estimates (P. Chen et al., 2012; Zheng, 2016), although adaptive design did show some improvement over random design under multidimensional IRT (P. Chen et al., 2017).
To truly optimize item calibration efficiency, the research question that should be answered is how to select an optimal sample of examinees so that the sample size needed to calibrate a set of new items can be reduced. This has been investigated from the perspective of optimal design theory (Berger, 1992, 1994; Buyske, 1998; Jones & Jin, 1994; van der Linden & Ren, 2015). The idea is similar to that of CAT, except that while CAT adaptively selects items for examinees to answer, in an optimal design study a select group of examinees (whose abilities are called design points in the optimal design literature) are assigned to the new items. The most common selection criteria in optimal design research are various optimality indices associated with the information matrix of the new item parameters. As in adaptive online calibration, the new item parameters are usually unknown and need to be initialized. A solution based on a sequential design was proposed by multiple authors (Berger, 1992, 1994; Y. I. Chang & Lu, 2010; Jones & Jin, 1994), and the general procedure is as follows. Starting with an initial set of item parameters, the system selects an optimal set of design points based on optimality indices and collects their responses. The item parameters are then updated from the responses and used for finding the optimal design points in the next iteration. This would be repeated until all new items are successfully calibrated.

The above procedures assumed the design points’ true values or estimates are somehow known, which is not practical in an online calibration scenario. Recently, a Bayesian optimal design approach (Ren et al., 2017; van der Linden & Ren, 2015) was proposed to address this issue. Instead of selecting design points for a given new item, they used the optimality indices to select new items for a given person. Unlike in CAT where the selected item has the largest amount of information at the current estimated
ability, the optimality indices would choose the item that would, among all the candidate new items, achieve the largest information gain when delivered to the current examinee. Their simulations, however, only compared between different optimality indices, and it was shown that their design actually performed very similarly to random selection in terms of calibration precision in the case of unidimensional (Zheng & Chang, 2017) and multidimensional CAT (P. Chen, 2017). An even more severe issue is that, as is pointed out by Zheng (2014) and demonstrated by their own simulation results (Ren et al., 2017), some new items would consistently register lower values on the optimality indices. This would lead to part of the new item pool being extremely underexposed and unable to get out of the calibration process when stopping rules are adopted. In fact, it was shown that as many as 40,000 examinees were required to calibrate just 100 new items simply because 20% of the new items were rarely administered (Ren et al., 2017).

Continuous online calibration. Online calibration is useful in calibrating new items as they are being developed and added over time, but none of the research mentioned above directly deals with the challenge of item bank development in the first place. There has been recent research that modified the procedures of online calibration into a strategy termed “continuous online calibration”, which allowed CAT to proceed without a full-fledged operational bank (Born, Fink, Spoden, & Frey, 2019; Fink, Born, Spoden, & Frey, 2018). Their approach is comparable to the adaptive online calibration mentioned above (P. Chen et al., 2017, 2012; Zheng, 2016) in that both have an initial phase in which new item parameter estimates are obtained. But instead of dividing a CAT session into multiple stages, they simulated multiple testing sessions and continuously calibrated part of the new item pool in each session. The reason this strategy can work
without a precalibrated item bank is that the very first session is simply a linear test where every examinee receives the same items. This provides preliminary item parameter estimates for use in the following sessions, which are each divided into an adaptive, calibration, and linking cluster. From the second session on, the design is essentially the same as traditional online calibration. The adaptive and linking cluster consist of items calibrated from previous sessions, and the calibration cluster contains the remaining new items that need to be estimated. After each session, the items from the calibration cluster are added to the operational pool, allowing for more adaptivity and higher quality link items in the following sessions.

In testing environments, obviously these calibration strategies would still lead to fairness issues because the adaptivity levels are not the same across different testing sessions. Given there is little adaptation at the beginning sessions (none in the first one, actually), those who take the test in these sessions would be at a big disadvantage as their scores would be much more unreliable and inaccurate than those taking the test later. For learning, this would be less of an issue because usually it is the same cohort of students that would go through multiple stages of assessment, and the scores from different cohorts or individuals are not compared with each other.

In summary, research in education and psychometrics offers some useful tools for online learning applications. AIG makes it possible to streamline, or even bypass, the item writing process by directly creating items from a small pool of existing items whose parameters are preferably known. Items made from AIG also allows for easier calibration later because the parent items already provide reasonably accurate initial parameter values. Online calibration has many design features that can be borrowed for use in the
learning context, but the estimation methods have the issue of not being truly online
(more details of the disadvantages of methods such as MEM will be discussed in Chapter 7).

**Ability tracking**

Because students are allowed to consult learning materials as well as provided feedback and/or hints while solving problems, it is expected that their ability would improve through practice. In psychometric terms, this means the ability level may change during the measurement process, and it is possible that ability change may occur after every item. In a standard CAT, however, it is assumed that examinee’s ability remains constant throughout the testing session. In order to accommodate the need to dynamically track ability changes on the fly, either the measurement model or the estimation algorithm must be modified.

**Longitudinal IRT models.** Within the IRT framework, there are some models developed to capture the longitudinal trajectory of ability levels. Earlier versions of longitudinal IRT models (Andersen, 1985; Embretson, 1991) are essentially multidimensional Rasch models in which each dimension represents a separate testing session. These models quickly become very complex as the number of assessment sessions increases, which is very common in the learning environment where students go through many practice sessions (Kadengye, Ceulemans, & Van den Noortgate, 2014). In addition, they cannot easily account for the changes occurring within a session.

Another dynamic extension of the Rasch model focuses on the learning process within a test (Verguts & De Boeck, 2000; Verhelst & Glas, 1993), and is similar to
Performance Factor Analysis in the field of educational data mining (Pavlik, Cen, & Koedinger, 2009). The exact formulation of the model depends on the assumption about the learning process, but the common feature of these models is that the response to current item not only depends on the item and person characteristics, but also the history of the student performance:

$$p_j(u_{ij} = 1|\theta_i) = \frac{e^{a_j(\theta_i + \lambda X_{ij} - b_j)}}{1 + e^{a_j(\theta_i + \lambda X_{ij} - b_j)}}$$ \hspace{1cm} (23)

Here $X_{ij}$ could be defined as the total number of correct responses by student $i$ up until item $j$ if it is assumed that learning only occurs when the student successfully solves an item. However, if learning is assumed to occur regardless of the actual response, then $X_{ij} \equiv j - 1$ and simply denotes the number of items attempted. Either way $X_{ij}$ changes as the student progresses through the session, giving the model its dynamic nature. $\lambda$ is the parameter that quantifies the learning effect, describing how much impact previous responses have on the current item. Note that $\lambda$ is considered to be person independent, which may not be a valid assumptions under many circumstances, and models that account for individual learning effects are still yet to be developed (Verguts & De Boeck, 2000). Parameter estimation can be done with either marginal maximum likelihood (Verhelst & Glas, 1993) or Gibbs sampling (Verguts & De Boeck, 2000). The model was shown to be useful in detecting whether learning has occurred and whether there are individual differences in learning effects using model comparisons (Verguts & De Boeck, 2000), but it has not been used in the assessment settings.

More recently, research on longitudinal IRT models adopted an explanatory IRT approach (de Boeck & Wilson, 2004) which, in most cases, uses time as the linear
predictor of ability. Furthermore, to account for the fact that growth trajectory varies across individuals, a random person-specific component is added to the intercept and slope of the linear growth model (Wang, Kohli, & Henn, 2016; Wang & Nydick, 2019). A longitudinal extension of Equation 3 is then given as

\[ p_j(u_{ij} = 1|\theta_i) = \Phi(\beta_j \theta_i^t + \alpha_j) \]  
\[ \theta_i^t = (\pi_0 + \omega_{0i}) + (\pi_1 + \omega_{1i}) * \text{time} + \epsilon_{it} \]

where \( \theta_i^t \) is now the ability of person \( i \) at time \( t \). Note that there is no superscript on the item parameters \( \alpha_j \) and \( \beta_j \), which means that the item parameters are assumed to the same over time. \( \pi_0 \) and \( \pi_1 \) are the group-level averages of intercept and slope, respectively. The random effects \( (\omega_{0i}, \omega_{1i}) \) are assumed to follow a bivariate normal distribution \( N(0, \Sigma) \) with \( \Sigma \) being the variance-covariance matrix to be estimated. The residual term \( \epsilon_{it} \) also follows a normal distribution with zero mean, and usually the variance is set to be the same across individual. Parameter estimates can be obtained using either maximum likelihood or weighted least squares (WLS). Although WLS leads to much faster computation, it may not work well in the learning environment because it cannot handle missing data properly (Wang et al., 2016), which is more likely to occur in online learning applications.

Variations of the longitudinal IRT model have been applied to the learning scenarios (Abbakumov, Desmet, & Van den Noortgate, 2019; Kadengye et al., 2014; Kadengye, Ceulemans, & Van Den Noortgate, 2015). These studies replaced the time variable in Equation 25 with those typically found in electronic learning, such as the number of sessions or the time spent within and between sessions, and fitted their models.
to the empirical data collected from online learning platforms. However, the models presented above all measure individual growth offline after data have been collected, which makes it incompatible with the goal of achieving adaptivity in item delivery on the fly. Due to the post-hoc nature of the longitudinal IRT models, many of the existing studies on adaptive learning systems made use of the Elo rating system (ERS; Elo, 1978) as the underlying model of adaptive learning systems (Pelánek, 2016).

Elo Rating System. ERS can be deemed as a special case of the paired-comparison models in the statistics literature (for a review, see Cattelan, 2012) and was originally proposed to measure the underlying ratings of chess players using results from chess games. The system is dynamic in the sense that players’ ratings can be updated after every match (or a set of matches, such as a tournament), making it an ideal choice for adaptive online learning where tracking the changes in learner ability on the fly is of primary concern.

The basic idea of the ERS is to calculate the expected probability of game outcome based on the current ratings of the two participating sides and update the ratings according to the difference between the expected probability and actual observed outcome. The winning player receives a higher rating after the update, and the update size depends on the rating difference between the two players. Consider the simplest scenario where there are no ties and a game between player $i$ and $j$ has only two possible outcomes: $u_{ij} = 1$ when player $i$ wins versus player $j$ and $u_{ij} = 0$ when player $i$ loses versus player $j$. Let $\theta_i$ and $\theta_j$ be the current ratings of both players and $p_{ij}$ the expected probability of player $i$ defeating player $j$, then both players’ new ratings after the game are given by
\[ \theta_i' = \theta_i + K(u_{ij} - p_{ij}) \]  
\[ \theta_j' = \theta_j - K(u_{ij} - p_{ij}) \]

The value $K$ is a weighting factor that can either be a constant or estimated from data (more details regarding values of $K$ will be given in the next chapter).

The expected probability $p_{ij}$ can be calculated in different ways, and just like in IRT, it typically follows the form of $p_{ij} = F(\theta_i - \theta_j)$ where $F$ is some S-shaped monotonically increasing function. The classic form adopted by the United States Chess Federation is based on the logistic function written as

\[ p_{ij} = \frac{1}{1 + 10^{\frac{\theta_i - \theta_j}{400}}} \]

while the normal ogive function is also a popular alternative. Generally, the choice of function $F$ does not have a substantial impact on model performance, and the logistic function seems to be favored due to its convenience in computation (Pelánek, 2016).

To adapt the system to educational applications, one can consider a chess game between two players as analogous to an interaction between an item and a learner. If the learner gets the item correct, this is considered a win for the learner whose ability level is higher than expected and would be increased as a result. Similarly, if an item is successfully solved by the learner, then the item may not be as difficult as expected and the difficulty parameter should be reduced by a certain amount. Due to this parallel between modeling chess games and learning outcome, the ERS can be easily adapted to the learning environment by rewriting $\theta_j$ in Equation 27 as the difficulty parameter $b_j$ of
item $j$. This also makes it possible in theory to estimate ability and item parameters at the same time. Note that when $p_{ij}$ is calculated with the standard logistic function, the ERS becomes identical in form to the Rasch model.

Currently, most of the research done on the topic of item-based adaptive learning systems was based on variations of the ERS model. The next chapter reviews these studies and also some of the limitations of using ERS in adaptive educational systems.

**Item selection**

In CAT, typically items with the largest information evaluated at the current ability level are selected to optimize measurement accuracy because the primary goal of testing is to estimate ability as precisely as possible. This is not always the case in the learning environment. For a learning system, the main purpose of the assessment is to facilitate the learning process, and in this case item selection methods commonly used in CAT may not be the optimal strategy.

**Learner motivation.** One important factor to consider for learning applications is learners’ motivation. When the 1PL or 2PL model is used in CAT, the system would typically administer items with a success probability of 50%. The value is slightly higher when the guessing parameter is included. It has been argued that this success rate was too low for learning and likely to result in reduced motivation and effort from the learners (Klinkenberg et al., 2011; Wauters et al., 2010). This is especially problematic for learning programs in which participation is not mandatory, as people will simply drop out when they feel discouraged and frustrated if they can only get about half of the items right even when trying their best. Therefore, item selection algorithms developed for the
learning environment should not solely consider psychometric properties, as learner motivation can also greatly influence the quality of responses.

Simulation has shown that for the 1PL model, increasing the desired success probability to as high as 70% only minimally affected ability estimates. The consequence of selecting easier items was more pronounced for 2PL models, resulting in much larger standard errors (Bergstrom, Lunz, & Gershon, 1992). A new item selection method was proposed to counter the negative effect of selecting overly easy items using traditional maximal information criterion (Eggen & Verschoor, 2006). The new method selects items with maximal information evaluated at a shifted ability level

\[
\hat{\theta} - \frac{1}{a_j} \log \frac{P_c}{1 - P_c}
\]  

where \(\hat{\theta}\) is the current ability estimate and \(P_c\) is the target success probability. Their results showed that the new method greatly reduced the loss in measurement precision when using easier or harder items, and it even worked well when exposure control was added.

Another potential approach of maintaining learner motivation is to allow the learners to choose, at least partially, the items and/or learning materials themselves (Wauters et al., 2010). For instance, the learners can simply ignore items that they believe are too easy or too difficult for them without any penalty, while in CAT these skipped items would oftentimes be scored as incorrect. In testing such a procedure is called a self-adapted test (SAT), in which items in the test are determined by the examinee instead of some item selection rule. SAT was mainly proposed as a measure to reduce test anxiety, and one meta-analysis (Pitkin & Vispoel, 2001) supported this claim by showing that
SAT resulted in lower post-test anxiety scores reported by examinees. They also found a slight increase in the ability estimates obtained from SAT, although the effect size was small and could be an artifact from the fact that estimates from SAT were more unreliable and biased. The authors of the meta-analysis argued that due to issues with test security and measurement accuracy, SAT should be used only in low-stakes assessment situations, which is exactly the case for many online learning applications. Therefore, adaptation of the SAT procedures to the learning environment could be a promising direction for future research.

**KL information.** Some researchers also discussed the use of KL information as the item selection criterion in the learning environment (Eggen, 2012; Veldkamp et al., 2011). As is mentioned in Chapter 2, KL information is a global measure of the distance between two probability distributions, and in the case of adaptive item selection, the distance between two likelihood functions evaluated at two ability levels. In learning, a natural choice of these likelihoods would be those evaluated at the learners’ current ability and the target ability specified by the instructor or the learner him/herself. In one simulation study, it was shown that when the ability levels increased during the test, item selection based on KL information outperformed Fisher information in terms of ability estimation, although the difference was only marginal (Eggen, 2012).

To summarize, due to the differences between learning and testing environments, many techniques of CAT are not directly applicable to the learning scenario. The major challenges can be summarized into three broad categories: item bank construction, dynamic ability estimation, and item selection. Some promising solutions have been proposed, but they are either not ideal for learning applications or need further research.
Before elaborating on the system based on the new Bayesian algorithm and how it is preferable to the existing solutions, the next chapter reviews the current state of research on adaptive learning systems.
Chapter 4: Current Research on Adaptive Learning Systems

As is mentioned in the previous chapter, much of the existing research on adaptive learning systems employed different variations of ERS to track learner ability on the fly. One of the first comprehensive studies on item-based adaptive learning systems was conducted by Klinkenberg et.al. (2011), in which they provided technical details on what they called the computerized adaptive practice (CAP) system. Their system was implemented in Math Garden, an online learning platform established to monitor and improve the math abilities of Dutch elementary school students. For the psychometric model, they proposed a modified version of ERS with the following changes. First, the update weight $K$ was used for both learner ability ($K_i$) and item difficulty ($K_j$), and allowed to vary as a function of the uncertainty $W$ in the parameter estimates.

\[
K_i = K_0 \left( 1 + K_+ W_i - K_- W_j \right)
\]

(30)

\[
K_j = K_0 \left( 1 + K_+ W_j - K_- W_i \right)
\]

(31)

Here $K_0 = 0.0075$, $K_+ = 4$, and $K_- = 0.5$ were all constants determined through pilot simulations. Each item and ability parameter has its own uncertainty parameter, and a new value of uncertainty $W'$ was calculated for both item and ability in the same way every time an item was administered:

\[
W' = W - \frac{1}{40} + \frac{1}{30} T
\]

(32)

$T$ is the number of days that elapsed since the last time the learner engages in the system. By definition, $W$ is bounded between 0 and 1 with a starting value of 1, and a larger $W$ value indicates greater uncertainty. This formula means that, in the case of learner ability,
the uncertainty would drop to 0 (and stay 0) after he/she finishes 40 consecutive items on the same day and would increase back to 1 after 30 days of inactivity.

Second, the scoring rules for both observed and expected outcomes were modified to incorporate response time data. The score received by the learner \( S_{ij} \) is now dependent on both the correctness of response \( u_{ij} \) (which is represented by 0 and 1, as usual) and the time \( t_{ij} \) needed by learner \( i \) to answer item \( j \)

\[
S_{ij} = (2u_{ij} - 1) \left(1 - \frac{t_{ij}}{y_j}\right) \tag{33}
\]

where \( y_j \) is the time limit specified for item \( j \). Therefore, the second term is a value between 0 and 1 and becomes larger when the response is fast. The first term is simply an indicator with 1 meaning the response is correct and 0 meaning incorrect. They named the scoring rule “high speed, high stakes” because if the response is correct, shorter response time would be more rewarding; but if incorrect, fast responses would lead to a big penalty. Therefore, quick responses always put the learners at a higher stake. The expected score was given as

\[
E(S_{ij}) = \frac{e^{2(\theta_i - b_j)} + 1}{e^{2(\theta_i - b_j)} - 1} - \frac{1}{(\theta_i - b_j)} \tag{34}
\]

The resulting ERS update rules were found by replacing \( u_{ij} \) and \( p_{ij} \) from Equation 26 and 27 with \( S_{ij} \) and \( E(S_{ij}) \)

\[
\theta_i' = \theta_i + K_i \left(S_{ij} - E(S_{ij})\right) \tag{35}
\]

\[
b_j' = b_j - K_j \left(S_{ij} - E(S_{ij})\right) \tag{36}
\]
Item selection was similar to the shifted ability approach proposed in Bergstrom et al. (1992). The desired success probability was chosen to be around 75%, which was achieved by selecting items with difficulty closest to

$$\hat{\theta} + \log \frac{p_c}{1 - p_c}$$

where $p_c$ was randomly drawn from $N(0.75,.1)$ then bounded between 0.5 and 1.

The authors claimed that the modified ERS could provide on-the-fly estimation of ability and difficulty parameters, although in their simulation study only the measurement precision of ability was reported. They compared the performance of the CAP system at varying levels of success probabilities with a standard CAT using Rasch model and weighted maximum likelihood estimators. The performance of the modified ERS that included individual’s response time data was similar to the standard CAT in terms of bias and standard error of ability estimates, even though the standard CAT approach did not have access to the information on individual response time. When only accuracy data were used, the original ERS performed much worse than standard CAT.

Several follow-up studies explored different aspects of the CAP system. Wauters, Desmet, and van den Noortgate (2012) investigated item parameter calibration and compared the item difficulty estimates derived from ERS with those from the Rasch model (using the marginal maximum likelihood estimator). It was concluded that ERS and the IRT estimation did not differ too much in terms of calibration accuracy with sample sizes varying from 20 to 300. Note that in this study the original version of ERS was used with $K$ fixed at 0.4, which was also decided from a pilot simulation study.
Another extension of the ERS-based learning system adopted the explanatory IRT modeling approach to provide better estimates on the starting values of learner ability by incorporating their previous learning history and background information (Park, Joo, Cornillie, van der Maas, & Van den Noortgate, 2019). As in Kadengye et al. (2015), they considered both the time spent within and between learning sessions as predictors of initial learner ability. In addition, individual learner characteristics such as age and gender could also be added. However, unlike Kadengye et al. (2015), the explanatory IRT model only served an assisting role by providing starting values at the beginning of each session, and once the system was initialized the ability estimate was still updated using traditional ERS. Their model was evaluated by multiple simulation studies in which the response data of 300 learners were generated, with 250 of them put into a training set where the model was fitted and model parameters such as π₀ and π₁ in Equation 25 were obtained. The remaining 50 learners constituted the validation set whose starting abilities could be calculated from the model fitted in the training set. The ability parameter recovery was only reported on the validation set. To simulate the online learning environment, instead of simulating a single practice session as in Klinkenberg et al. (2011), there were a total of 4 sessions with either 40 or 80 items in each session, and both within-session and between-session change were simulated. The update weight K was set to be 0.7 in the beginning and linearly decreased as more items were answered within each session. The results showed an improvement in ability estimates achieved by adding the explanatory variables, especially during the early period of a session.

More recently, a multidimensional version of ERS in CAP environments was proposed in an attempt to monitor multiple abilities simultaneously (Park, Cornillie, van
der Maas, & Van Den Noortgate, 2019). The parameter update rules were similar to those in Klinkenberg et al. (2011) and given as follows:

\[
\theta'_{im} = \theta_{im} + D_{jm} K_i \left( S_{ij} - E(S_{ij}) \right) \tag{37}
\]

\[
b'_j = b_j - D_{jm} K_j \left( S_{ij} - E(S_{ij}) \right) \tag{38}
\]

Now a total of \(M\) dimensions of abilities were being measured and \(D_{jm}\) was an indicator variable on whether the current item measures ability \(m\). Including the indicator variable \(D_{jm}\) requires that the loading structure of all items needs to be known beforehand, and as a result the item bank was assumed to be precalibrated for simplicity. The update weight \(K\) was also set to be monotonically decreasing as more items were answered and bounded between 0.4 and 0.1. A single session of 200 items was simulated and answered by a sample of 250 learners, and ability was updated on the fly after each response. Ability estimation became more accurate as more items were answered by a student, and the more items there were for a specific dimension, the better the recovery for that dimension.

The ERS-based approach provided practically feasible solutions to many of the challenges encountered in building an adaptive learning system. According to Pelanek (2016), the system “is particularly attractive when we want to build a reasonably behaving system quickly and cheaply” (p.177). The system can function without knowledge of any of the item parameters, so the costly calibration studies are no longer needed. It is also fast to compute, which makes it suitable for the tasks of real-time parameter estimation and dynamic ability tracking. All these benefits, however, come at the cost of compromised measurement accuracy, due to the simplistic nature of the
ERS is mathematically similar to the Rasch model which is the simplest IRT model that does not allow items to vary in discriminating power. But the bigger issue is the arbitrariness in setting the update weight $K$. While the Rasch model uses conditional maximum likelihood to estimate its parameters, ERS does it through an update rule that relies heavily on the choice of $K$, which directly determines the update size. For ability levels, a large value of $K$ means that the estimates would constantly fluctuate without converging to a certain point, while a small value means it would take too many items to close in on the true value (assuming the ability is constant over time) or to catch up to the ability change that is taking place. Ideally $K$ should be estimated from data as well, but in the literature reviewed above, this was not the case as $K$ was either fixed before data collection or computed by some linear equations that included many tuning parameters chosen based on pilot simulation or previous research. As the optimal values of $K$ are likely to vary across different situations, there is no way of knowing whether the choice of $K$ is appropriate until some post-hoc analysis is done. As a result, the estimation methods in such systems are much less based on statistical grounds than those used in IRT research, and the resulting estimates are not as accurate and trustworthy in comparison. This could also partially be the reason that, although ERS has the potential to estimate item parameters at the same time as ability, in simulations discussed so far, either precalibrated item banks were assumed as in standard CAT, or the calibration results were never reported. Therefore, under the circumstances where the item bank was not precalibrated, the ERS-based adaptive learning systems probably could not achieve the intended level of adaptivity due to the item parameters being poorly estimated.
In the next chapter, an improved update algorithm under the Bayesian framework is proposed. While keeping all the advantages of ERS, the new algorithm includes a variance component for each parameter to quantify the amount of uncertainty around its estimate, and the update size is now data-driven and computed in real time using the current responses and parameter estimates. The new algorithm can also estimate item discrimination so that now each item could have its own discrimination parameter, which is more flexible than ERS that assumes equal discrimination for all items.
Chapter 5: Introduction to the New Bayesian Update Rules

Theoretical Background

The new update rules are based on the work by Weng and Coad (2018), which, similar to ERS, stems from the research on the analysis of dynamic paired comparison data in the statistics literature. In the original ERS that was widely used in chess rating, the abilities were assumed to be point estimates. Bayesian extension of this model was later developed to incorporate the precision or variability of the ability estimates as well (Fahrmeir & Tutz, 1994; Glickman, 1993). In this formulation, ability was normally distributed with time-varying means and variance. Instead of being fixed a priori, the update weight $K$ was dynamically calculated from the observed outcome and the previous parameter estimates and their variances. The precision of an ability estimate affects the ability update in the following way. If the ability estimate contains substantial uncertainty (for example, a new player has just entered the rating system or there is a long interval since the player’s previous activities), the next outcome would have a big impact on the ability update. When there is little certainty about the ability level, then outcomes would not affect the rating too much (a player with a long and mostly winning record should not suddenly drop too many points after a few unexpected losses).

These new Bayesian models were estimated offline with either MCMC (Glickman, 1993) or empirical Bayes (Fahrmeir & Tutz, 1994), and they become unwieldy when the number of players or outcomes gets large. A non-iterative algorithm with closed-form equations for the posterior means and variance that allows for online data processing was derived by approximating the posterior distributions of ability
The algorithm was the basis of the Glicko rating system that has been used extensively in chess and other competitive online games. Further extensions of the Bayesian paired comparison models include the TrueSkill rating system developed by Microsoft for their Xbox™ consoles that could model data from matches of more than two players (Herbrich, Minka, & Graepel, 2007). An improved version called TrueSkill 2 was proposed more recently to incorporate collateral information about the players in the game (Minka, Cleven, & Zaykov, 2018). Although TrueSkill utilized a deterministic algorithm to approximate the posterior distribution, it is still an iterative procedure and leaves room for further simplification like the closed-form computational formulas given in Glickman (1999).

Such formulas were provided in Weng and Lin (2011) for paired comparison models and Weng and Coad (2018) for IRT models. The general steps of their derivation are summarized below. Suppose $\xi_i$ is a model parameter with prior distribution of $N(\mu_i, \sigma_i^2)$ and let

$$z_i = \frac{\xi_i - \mu_i}{\sigma_i}$$  \hspace{1cm} (39)$$

The posterior distribution of the parameter vector $\mathbf{z} = (z_1, z_2, \ldots, z_i, \ldots, z_p)^T$ can be written as

$$f(\mathbf{z}|\mathbf{u}) = \frac{\phi(\mathbf{z})f(\mathbf{u}|\mathbf{z})}{\int \phi(\mathbf{z})f(\mathbf{u}|\mathbf{z})d\mathbf{z}},$$  \hspace{1cm} (40)$$

where $f(\mathbf{u}|\mathbf{z})$ is the likelihood function of data vector $\mathbf{u}$ and $\phi$ denotes the density of a $p$-variate standard normal variable. Weng and Lin derived a corollary based on a lemma by Woodroofe (1989), which was concerned with computing the expectation of a “nearly
normal” distribution of the form \( \phi(z)f(u|z) \). Using the corollary, if the posterior distribution of \( z \) can be expressed in the form of Equation 40, then the first and second moment of the posterior distribution of \( z \), assuming \( f(u|z) \) is twice differentiable, can be found as

\[
E(z|u) = E \left( \frac{\nabla f(u|z)}{f(u|z)} \right) = \mu \tag{41}
\]

\[
E(z_i^2|u) = 1 + E \left( \frac{\nabla^2 f(u|z)}{f(u|z)} \right), \quad i = 1, \ldots, p \tag{42}
\]

where by definition \( \nabla f(u|z) \equiv \frac{\partial f(u|z)}{\partial z} \) and \( \nabla^2 f(u|z) \equiv \frac{\partial^2 f(u|z)}{\partial z^2} \).

The first moment can be further simplified by applying the chain rule to the first derivative in Equation 41. For a single parameter \( z_i \)

\[
E(z_i|u) = E \left( \frac{\partial f(u|z)}{\partial z_i} \right) = E \left( \frac{\partial \log f(u|z)}{\partial z_i} \right) \tag{43}
\]

The variance can be found as

\[
\text{var}(z_i|u) = E(z_i^2|u) - E(z_i|u)^2 = 1 + E \left( \frac{\nabla^2 f(u|z)}{f(u|z)} \right) - E \left( \frac{\partial \log f(u|z)}{\partial z_i} \right)^2 \tag{44}
\]

Now based on Equation 39, the posterior means and variance of \( \xi_i \) are

\[
\mu_i' = E(\xi_i|u) = \mu_i + \sigma_i E(z_i|u) = \mu_i + \sigma_i E \left( \frac{\partial \log f(u|z)}{\partial z_i} \right) \tag{45}
\]

\[
\sigma_i'^2 = \text{var}(\xi_i|u) = \sigma_i^2 \text{var}(z_i|u) = \sigma_i^2 \left( 1 + E \left( \frac{\nabla^2 f(u|z_i)}{f(u|z_i)} \right) - E \left( \frac{\partial \log f(u|z)}{\partial z_i} \right)^2 \right) \tag{46}
\]
Now apply the above results to the GRM model described in Equation 7. Assuming person and item parameters follow normal priors and are mutually independent, it is possible to derive a series of equations to update the means and variance that define their posterior distributions. The primary challenge here is to evaluate the expectations on the right-hand sides of the Equations 45 and 46, and some sort of approximation is necessary for non-iterative updater rules that are simple to compute. In the derivation given in Weng and Coad (2011), the most important idea is to replace the random variables with their corresponding means. To be more specific, let the prior distributions be $\alpha_j \sim N\left(\mu_{\alpha_j}, \sigma^2_{\alpha_j}\right)$, $\beta_j \sim N\left(\mu_{\beta_j}, \sigma^2_{\beta_j}\right)$, and $\theta_i \sim N\left(\mu_{\theta_i}, \sigma^2_{\theta_i}\right)$. Also let $z = (\alpha_j^*, \beta_j^*, \theta_i^*)^T$ where $\alpha_j^* = \frac{\alpha_j - \mu_{\alpha_j}}{\sigma_{\alpha_j}}$, $\beta_j^* = \frac{\beta_j - \mu_{\beta_j}}{\sigma_{\beta_j}}$, and $\theta_i^* = \frac{\theta_i - \mu_{\theta_i}}{\sigma_{\theta_i}}$. For a single response, the likelihood function, according to Equation 7, is

$$f(u|z) = \Phi(\beta_j \theta_i + \alpha_j - \gamma_c) - \Phi(\beta_j \theta_i + \alpha_j - \gamma_{c+1})$$

(47)

Take $\alpha_j$ as an example, the derivatives in the expectations of the equations 45 and 46 are, applying the chain rule,

$$\frac{\partial \log f(u|z)}{\partial \alpha_j^*} = \sigma_{\alpha_j} \Omega(x_1, x_2)$$

(48)

$$\frac{\nabla^2 f(u|\alpha_j^*)}{f(u|\alpha_j^*)} = \sigma^2_{\alpha_j} \left(\left(\Omega(x_1, x_2)\right)^2 - \Delta(x_1, x_2)\right)$$

(49)

where

$$x_1 = \beta_j \theta_i + \alpha_j - \gamma_c$$

(50)

$$x_2 = \beta_j \theta_i + \alpha_j - \gamma_{c+1}$$

(51)
\[
\Omega(x_1, x_2) = \frac{\phi(x_1) - \phi(x_2)}{\Phi(x_1) - \Phi(x_2)}
\]  

(52)

\[
\Delta(x_1, x_2) = \frac{x_1\phi(x_1) - x_2\phi(x_2)}{\Phi(x_1) - \Phi(x_2)} + \Omega(x_1, x_2)^2
\]  

(53)

Then

\[
E\left( \frac{\partial \log f(u|z)}{\partial \alpha_j} \right) = \sigma_{\alpha_j} E(\Omega(x_1, x_2))
\]  

(54)

Expand the expectation on the right as a triple integral

\[
E(\Omega(x_1, x_2))
= \int \int \int \Omega(x_1, x_2) \phi(\alpha_j|\mu_{\alpha_j}, \sigma_{\alpha_j}^2) \phi(\beta_j|\mu_{\beta_j}, \sigma_{\beta_j}^2) \phi(\theta_i|\mu_{\theta_i}, \sigma_{\theta_i}^2) d\alpha_j d\beta_j d\theta_i
\]

\[
= \int \int \int \frac{[\phi(x_1) - \phi(x_2)]}{[\Phi(x_1) - \Phi(x_2)]} \phi(\alpha_j|\mu_{\alpha_j}, \sigma_{\alpha_j}^2) \phi(\beta_j|\mu_{\beta_j}, \sigma_{\beta_j}^2) \phi(\theta_i|\mu_{\theta_i}, \sigma_{\theta_i}^2) d\alpha_j d\beta_j d\theta_i
\]

where \(\phi(\cdot)\) is the normal density with given mean and variance. The first step in simplifying the integral is to apply the approximation \(E(X^r) \approx \frac{E(x^r)}{E(Y)}\) so that

\[
E(\Omega(x_1, x_2)) \approx \int \int \int \frac{[\phi(x_1) - \phi(x_2)]}{[\Phi(x_1) - \Phi(x_2)]} \phi(\alpha_j|\mu_{\alpha_j}, \sigma_{\alpha_j}^2) \phi(\beta_j|\mu_{\beta_j}, \sigma_{\beta_j}^2) \phi(\theta_i|\mu_{\theta_i}, \sigma_{\theta_i}^2) d\alpha_j d\beta_j d\theta_i
\]

Now the problem can then be broken down into simplifying individual integral such as

\[
\int \int \phi(x_1) \phi(\alpha_j|\mu_{\alpha_j}, \sigma_{\alpha_j}^2) \phi(\beta_j|\mu_{\beta_j}, \sigma_{\beta_j}^2) \phi(\theta_i|\mu_{\theta_i}, \sigma_{\theta_i}^2) d\alpha_j d\beta_j d\theta_i
\]

1 The approximation in this step makes use of a first-order Tyler series expansion and is most accurate when the variances of \(X\) and \(Y\) are small. For a more detailed proof see https://www.stat.cmu.edu/~hseltman/files/ratio.pdf.
which is actually the integral of the product of 4 normal densities. It can be easily shown that the identity

\[ \int \phi(x|\mu_x, \sigma_x^2)\phi(\mu_x|\mu_0, \sigma_0^2)d\mu_x = \phi(x|\mu_0, \sigma_x^2 + \sigma_0^2) \]

holds. Write \( \phi(x_1) = \phi(\beta_j \theta_i + \alpha_j - \gamma_c) = \phi(\gamma_c - \beta_j \theta_i|\alpha_j, 1) \) and apply the above identity to simplify the integral. For example,

\[ \int \phi(\gamma_c - \beta_j \theta_i|\alpha_j, 1)\phi(\alpha_j|\mu_{\alpha_j}, \sigma_{\alpha_j}^2)d\alpha_j = \phi(\lambda_j - \beta_j \theta_i|\mu_{\alpha_j}, 1 + \sigma_{\alpha_j}^2) \]

The identity can be applied three times, and for the last time it is necessary to replace \( \theta_i \) with \( \mu_{\theta_i} \) in the variance component of the normal density. The result of the simplification is a new normal PDF

\[ \phi(\gamma_c - \mu_{\alpha_j}, \mu_{\beta_j}, \mu_{\theta_i} - \gamma_c, 1 + \sigma_{\alpha_j}^2 + \sigma_{\beta_j}^2 + \sigma_{\theta_i}^2) \]

Apply similar procedures to the denominator (there is a similar identity for the product of a cumulative normal density as well) and replace \( \theta_i \) with \( \mu_{\theta_i} \) when needed, the expectation can be approximated as

\[ E(\Omega(x_1, x_2)) \approx \frac{1}{v} \phi(x_1^*) - \phi(x_2^*) = \frac{1}{v} \Omega(x_1^*, x_2^*) \]

where
\[ v = \sqrt{1 + \sigma^2_{\alpha_j} + \sigma^2_{\beta_j} \mu^2_{\theta_i} + \sigma^2_{\theta_i} \mu^2_{\beta_j}} \]  

(56)

\[ x_1^* = \frac{\mu_{\alpha_j} + \mu_{\beta_j} \mu_{\theta_i} - Y_c}{v} \]  

(57)

\[ x_2^* = \frac{\mu_{\alpha_j} + \mu_{\beta_j} \mu_{\theta_i} - Y_{c+1}}{v} \]  

(58)

The expectations for \( \beta_j \) and \( \theta_i \) can be found analogously, and the update rules for the posterior means and variances of item and ability parameters are given below

\[ \mu'_{\alpha_j} = \mu_{\alpha_j} + \frac{\sigma^2_{\alpha_j}}{v} \Omega(x_1^*, x_2^*) \]  

(59)

\[ \mu'_{\beta_j} = \mu_{\beta_j} + \frac{\mu_{\theta_i} \sigma^2_{\beta_j}}{v} \Omega(x_1^*, x_2^*) \]  

(60)

\[ \mu'_{\theta_i} = \mu_{\theta_i} + \frac{\mu_{\beta_j} \sigma^2_{\theta_i}}{v} \Omega(x_1^*, x_2^*) \]  

(61)

\[ \sigma^2_{\alpha_j}' = \sigma^2_{\alpha_j} \max \left( 1 - \left( \frac{\sigma_{\alpha_j}}{v} \right)^2 \Delta(x_1^*, x_2^*), \kappa \right) \]  

(62)

\[ \sigma^2_{\beta_j}' = \sigma^2_{\beta_j} \max \left( 1 - \left( \frac{\mu_{\theta_i} \sigma_{\beta_j}}{v} \right)^2 \Delta(x_1^*, x_2^*), \kappa \right) \]  

(63)

\[ \sigma^2_{\theta_i}' = \sigma^2_{\theta_i} \max \left( 1 - \left( \frac{\mu_{\beta_j} \sigma_{\theta_i}}{v} \right)^2 \Delta(x_1^*, x_2^*), \kappa \right) \]  

(64)

where \( \kappa \) is a constant slightly larger than 0 to make sure that posterior variance is always positive.

These equations are complex in form, but it is still possible to make some intuitive sense of them. For instance, the posterior mean of \( \theta \) depends on the prior mean
of the discrimination parameter so that an item with a large discrimination would have a larger effect on $\theta$ update. Also note that the size of parameter means update has a positive relationship with the parameter variance. Because parameter variance is monotonically decreasing (every update on variance is essentially multiplying the previous value by a number smaller than 1), it is expected that the parameters will fluctuate the most at the early stage and gradually stabilizes as more data accumulate.

The new algorithm is intended as an improvement over the existing adaptive learning system based on ERS. It provides encouraging solutions to the issues of dynamic ability estimation and item bank construction by providing on-the-fly estimation for both item and person parameters. With regard to item selection, the new algorithm can be readily combined with aforementioned item selection rules, such as maximal information at shifted ability, to facilitate the learning process, although in this dissertation only the standard CAT item selection methods will be considered.

**Practical Implementation**

Implementing the Bayesian update rules should be as straightforward as the ERS-based systems, but some special considerations are outlined below.

**Order of Data Entry.** This is an issue unique to the system based on the new parameter update rules because item and person parameters are now computed at the same time, thus affecting each other. In IRT, parameter estimation proceeds in a two-step fashion, with item parameters first estimated with MMLE/EM and ability computed next based on the estimated item parameters. An undesirable consequence of this simultaneous
estimation scheme is that it will not function as intended unless the response are entered into the system in a random order.

To see why, first it is worth identifying the assessment scenarios that vary in terms of the orders in which data are entered. For online learning platforms, as well as online surveys, people do not start responding at the same time and in many cases, data enter the system sequentially as a learner only begins after the previous person has finished the session. This will be named the by-person design and illustrated in Figure 5.1. Learner $i$ completes $L$ items, and by the time he/she is finished, the next learner starts his/her own set of items.

In standardized testing, however, usually examinees start the test at the same time and go through the test at their own pace. In the extreme case where all examinees spend the same time on each item, data entry can be conceptualized as following a by-item design (Figure 5.2) in which the next item is delivered only after the previous item has
been answered by every examinee. Note that here the item numbering simply refers to the item position on each person’s test, and an item $j$ may (in linear testing) or may not (in adaptive testing) be the same item for different examinees.

**Figure 5.2: By-item design, where an item is presented after every learner has finished the previous item**

The by-person and by-item design only represent the extreme scenarios, while in practice the data are likely to come in following a mixed pattern. These designs are useful, however, in demonstrating the effect of data entry order on the new parameter update rules. Consider the scenario where the learners enter the learning system under the by-person scheme and those who enter the system first come from a limited range of ability. As is pointed out earlier, parameter changes slow down as posterior variances become smaller. This becomes problematic when parameter estimates are far from their true values after the first few updates. Because the update size gradually decreases, it would be difficult for the estimate to revert back to the true value. Suppose the order in which learners enter the session is completely determined by their true abilities from highest to lowest. In the beginning, the set of items administered to this group of learners
would appear easier than they actually are because they are only answered by the high-ability students. These inaccurate estimates are unlikely to be corrected by response data that come in later due to the shrinking update size. In addition, these faulty parameter estimates would in turn “contaminate” the ability estimates of the learners who receive them later, and the errors would then probably propagate to most of the parameter estimates in the system. This is indeed confirmed by one of the pilot simulations in which large bias in both person and ability parameter was observed when the order of entry was determined by learner ability for both linear and adaptive testing. The large bias disappeared when the data were entered randomly and was reduced under the by-item design for adaptive testing as not all items were necessarily first delivered to learners of a certain ability range (there was no difference between both designs for linear testing when the order of entry was non-random because all items were still first answered by learners from a certain range of ability).

The second aspect to the issue is that under by-person design the early cohort of learners usually end up having larger posterior variance, regardless of their ability levels, in part because the items given to them are barely calibrated and have large posterior variance as well. By-item design on the other hand, could help eliminate this effect because the early persons are able to receive some well-calibrated items at the end of their tests, while in the by-person condition all items answered by the early persons were minimally calibrated. This was again confirmed in pilot simulations, which also showed that there was a modest correlation between posterior variance and the order in which learners entered the session. This put the early cohort at a disadvantage and is problematic in educational testing where fair scoring of individuals is of high priority. In
learning environment, this becomes less of an issue because learners usually engage in more than one session, and as long as the order is not identical for all sessions, this ordering effect could be cancelled out over multiple sessions.

In conclusion, the order in which learners’ data enter the assessment has a substantial impact on the performance of the new update rules. It is not unreasonable to say that in most applications of online learning, learners do not start their session at the same time, making it closer to a by-person design. Fortunately, the negative influence of the order effect can be largely avoided with multiple rounds of assessment, which is typical of online learning systems. Note that sometimes it is possible that the order of entering a session is indeed related to learner ability (for instance, it is possible that higher-achieving learners are also more motivated and less likely to procrastinate) and this relationship would persist through multiple sessions. Therefore, it is advisable to carefully collect and monitor data on user activities and take necessary precautions to combat the order effect (for example, setting a small time frame when the session will be available, thus forcing a by-item design).

**Update Rules for the Binary IRT Model.** The update equations presented in Equation 59 to 64 are based on the GRM in Equation 47 as a more general form. As is shown in Chapter 2, GRM can be reduced to binary model by letting $C = 2$ and $γ_1 = 0$. As a result Equation 47 becomes $P(u_{ij} = 1 | α_j, β_j, θ_i) = Φ(β_jθ_i + α_j)$ when $c = 1$ (correct answer) and $P(u_{ij} = 0 | α_j, β_j, θ_i) = 1 − Φ(β_jθ_i + α_j)$ when $c = 0$ (incorrect answer). The update equations can thus be simplified accordingly.
Let the prior means and variance of the ability parameter be $\mu_{\theta_i}^{(0)}$ and $\sigma_{\theta_i}^2(0)$, respectively. After responding to $m$ items, the current parameter estimates become $\mu_{\theta_i}^{(m)}$ and $\sigma_{\theta_i}^2(0)$. Suppose the response to the $(m + 1)th$ item is correct, then the next parameter estimates are given as

$$
\mu_{\theta_i}^{(m+1)} = \mu_{\theta_i}^{(m)} + \mu_{\beta_j}^{(m)} \sigma_{\theta_i}^2(0) \frac{\phi(x^{*}(m))}{\Phi(x^{*}(m))}
$$

(65)

$$
\sigma_{\theta_i}^2(m+1) = \sigma_{\theta_i}^2(m) \max \left( 1 - \left( \frac{\mu_{\beta_j}^{(m)} \sigma_{\theta_i}^2(0)}{\nu(m)} \right)^2 \left[ \frac{x^{*}(m)}{\Phi(x^{*}(m))} + \frac{\phi(x^{*}(m)^2)}{\Phi(x^{*}(m))^2} \right], \kappa \right)
$$

(66)

where $x^{*}(m) = \frac{\mu_{\theta_i}^{(m)} + \mu_{\beta_j}^{(m)} \mu_{\theta_i}^{(m)}}{\nu(m)}$. If the $(m + 1)th$ response is incorrect, the equations are

$$
\mu_{\theta_i}^{(m+1)} = \mu_{\theta_i}^{(m)} - \mu_{\beta_j}^{(m)} \sigma_{\theta_i}^2(0) \frac{\phi(x^{*}(m))}{1 - \Phi(x^{*}(m))}
$$

(67)

$$
\sigma_{\theta_i}^2(m+1) = \sigma_{\theta_i}^2(m) \max \left( 1 - \left( \frac{\mu_{\beta_j}^{(m)} \sigma_{\theta_i}^2(0)}{\nu(m)} \right)^2 \left[ \frac{-x^{*}(m) \phi(x^{*}(m))}{1 - \Phi(x^{*}(m))} + \frac{\phi(x^{*}(m)^2)}{[1 - \Phi(x^{*}(m))]^2} \right], \kappa \right)
$$

(68)

Based on Equations 65 and 67, it is evident that the posterior mean of the ability parameter goes up after a correct response (plus an always positive value since $\beta_j$ is always positive and $\Phi$ is a CDF) and goes down after an incorrect response, which makes intuitive sense and also agrees with the ERS formulas. Another noteworthy observation from the equations is that the next updated estimate only depends on the response data and the parameter estimates from the previous cycle. This is analogous to the
“memoryless” property of the Markov chains and means that for the purpose of tracking learner ability, it is possible in practice to only extract the most current state of the system rather than the learner’s entire response history. This is in contrast to CAT and online calibration in which learner ability has to be estimated using the response vector and parameters of all the previously answered (operational) items.

The update formulas of the item parameters for binary models can be obtained similarly. Take the intercept parameter as an example. For the correct and incorrect response, the intercept parameter can be updated as follows:

\[
\mu_{\alpha_i}^{(m+1)} = \mu_{\alpha_i}^{(m)} + \sigma_{\alpha_i}^2(m) \frac{\phi(x^{*}(m))}{\phi(x^{*}(m))}, \tag{69}
\]

\[
\sigma_{\alpha_i}^2(m+1) = \sigma_{\alpha_i}^2(m) \max \left( 1 - \left( \frac{\sigma_{\alpha_j}^2(m)}{\nu(m)} \right)^2 \left[ \frac{x^{*}(m) \phi(x^{*}(m))}{\phi(x^{*}(m))} + \frac{\phi(x^{*}(m))^2}{\phi(x^{*}(m))^2} \right], k \right), \tag{70}
\]

\[
\mu_{\alpha_i}^{(m+1)} = \mu_{\alpha_i}^{(m)} - \frac{\sigma_{\alpha_j}^2(m)}{\nu(m)} \frac{\phi(x^{*}(m))}{1 - \Phi(x^{*}(m))}, \tag{71}
\]

\[
\sigma_{\alpha_i}^2(m+1) = \sigma_{\alpha_i}^2(m) \max \left( 1 - \left( \frac{\sigma_{\alpha_j}^2(m)}{\nu(m)} \right)^2 \left[ \frac{-x^{*}(m) \phi(x^{*}(m))}{1 - \Phi(x^{*}(m))} + \frac{\phi(x^{*}(m))^2}{[1 - \Phi(x^{*}(m))]^2} \right], k \right). \tag{72}
\]

Here the intercept also increases with a correct response, but because the intercept is in the opposite direction of item difficulty \((\alpha_j = -a_jb_j)\), the difficulty parameter would actually drop given a correct response, which intuitively means the item is not as difficult as previously thought and is again consistent with the ERS formulas.
The process of parameter update is further illustrated by Figure 5.3, which also highlights the “memoryless” property of the new update rules. Parameter estimates are performed each time a learner responds to an item. After the response is recorded and scored, only the most current parameter estimates of the learner and the item are fed into a set of equations, depending on whether the response was correct or not, from which the updated parameter estimates are to be calculated for both the learner and the item. These new parameter estimates replace the previous ones and will be used in the next update when needed, and the process continues until the entire session is finished.

**Figure 5.3 Flowchart of Parameter Update Using the Bayesian Algorithm**
Chapter 6: Simulation studies

This chapter presents two simulation studies that evaluate the performance of the new algorithm as an assessment tool in the adaptive learning environment. The first study was implemented in a setting similar to the traditional testing environment with only one session and was presented as evidence that the new algorithm was comparable to the IRT methods in terms of psychometric properties. The second study simulated a more realistic online learning scenario and provided more direct support for the algorithm as a usable and useful technique in the computerized learning environment. All simulations and procedures were performed in R 3.6.1 (R Core Team, 2019) and the relevant code was included in the Appendix.

Study 1

The purpose of the first study was to compare the performance of the new algorithm with MMLE/EM used for IRT in terms of parameter recovery under different specifications of linear testing. In addition, since the algorithm is built upon a few assumptions about the measurement model and parameter distributions, a robustness analysis was conducted to see if the algorithm would still work given violations of these assumptions.

Simulation design. In the main simulation, a single assessment session of a fixed length was simulated. 5 levels of test length (L=10, L=20, L=40, L=100, and L=200, with the shorter tests not nested within the longer ones) and 4 levels of sample size (N=200, N=500, N=1000, and N=2000) were fully crossed to produce 20 conditions. Following Weng and Coad (2018), the intercept and ability parameters were generated from
\( \alpha_j \sim \text{N}(0, 1^2) \) and \( \theta_i \sim \text{N}(0, 1^2) \), respectively. In their simulation, they generated \( \beta_j \) from \( \text{N}(1, 20) \) because they were evaluating the performance of the algorithm for online rating data, which, unlike in educational settings, allowed for negative discrimination (this means that a rater would assign a more negative rating to a product positively reviewed by others, which may or may not be a problem). This is usually not appropriate in the learning situation where negative item discrimination is to be avoided. In this study the slope parameter was instead generated from \( \beta_j \sim \text{N}(1, 0.32^2) \) and the prior distribution was \( \beta_j \sim \text{N}(1, 0.6^2) \). To guarantee there were no negative slope parameters, each set of item parameters was checked to make sure all of the slope parameters were larger than 0.1 and would have to be regenerated until this condition is satisfied. These item parameter distributions were comparable to those used in previous simulations (Patton, Cheng, Yuan, & Diao, 2013; Ren et al., 2017), which extracted their items from an existing operational item pool.

Responses were generated from the 2-parameter binary normal ogive model using true item and ability parameters, and a set of parameter estimates was obtained with the Bayesian update rules after every new response. The data followed the by-person design with the learners entering the system randomly so that the undesirable order effect would not be of primary concern. With regard to the traditional item calibration methods, item parameters were obtained by applying the MMLE/EM algorithm once on the entire response matrix after the session was finished, and \( \theta \) estimates were computed based on the estimated item parameters using Expected A Priori (EAP) estimation with standard normal prior.
For the robustness analysis, 3 different kinds of assumption were simulated, and the comparison was also made between the Bayesian update formulas and MMLE/EM. First, response data were generated with the 3-parameter model, with a pseudo-guessing parameter of 0.20, then fitted with the 2-parameter models to see how model misspecification would affect the measurement performance. Second, the data-generating distribution for the slope parameter was changed to a uniform distribution \( U(0.5, 2) \) while the prior distribution remained unchanged. Third, the data-generating distribution for the ability parameter was changed to skewed-normal with the same mean and standard deviation but a skewness of 0.9 while the prior remained standard normal. The latter two conditions were intended as a test of how sensitive the Bayesian update rules were to the violations of distributional assumptions. The robustness analysis did not include all the conditions from the main simulation and only followed a 2 (L=40 and L=100) by 2 (N=500 and N=2000) crossed design, while all the remaining settings and procedures remained the same.

Item parameter recovery for the main simulation and the robustness analysis was evaluated by the bias and root mean square error (RMSE) averaged across 30 replications for each item. The overall bias and RMSE for the slope and intercept parameter were then averaged across all the items. For \( \theta \), the bias and RMSE was first calculated within each replication, then an average was taken on the 30 bias and RMSE values as the measure of \( \theta \) recovery.

\[
bias(\alpha_j) = \frac{1}{30} \sum_{r=1}^{30} (\hat{\alpha}_j - \alpha_j)
\]

(73)
\[
RMSE(\alpha_j) = \sqrt{\frac{1}{30} \sum_{r=1}^{30} (\hat{\alpha}_j - \alpha_j)^2} \quad (74)
\]

\[
bias(\theta) = \frac{1}{30} \times \frac{1}{N} \sum_{i=1}^{N} (\hat{\theta}_i - \theta_i) \quad (75)
\]

\[
RMSE(\theta) = \frac{1}{30} \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{\theta}_i - \theta_i)^2} \quad (76)
\]

**Results.** Table 6.1 shows the parameter recovery for both the new algorithm and the traditional MMLE/EM and EAP method from the main simulation. The Bayesian algorithm computed much faster than MMLE/EM, especially in conditions with large response matrix. In addition, in the L=100 and N=200 condition, there were two datasets that failed to converge, and another two datasets had an item with perfect responses, hence the results were only based on the remaining 26 datasets. For \( \theta \) recovery, EAP consistently registered smaller RMSE than the new update rules. The new method was generally positively biased, especially when the test was long, while EAP yielded minimal bias in all conditions. Both methods gained substantially from increased test length and much less from increased sample size. The difference in RMSE was largely the same across different test length/sample size combinations, mostly in the range of 0.10 and 0.30. In sum, the new update algorithm behaved very similarly to EAP, with an acceptable loss in precision due to the numerical approximation involved.
<table>
<thead>
<tr>
<th>Test length</th>
<th>Sample size</th>
<th>( \alpha )</th>
<th></th>
<th>( \beta )</th>
<th></th>
<th>( \theta )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Bias &amp; RMSE</td>
<td>Bias &amp; RMSE</td>
<td>Bias &amp; RMSE</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>New EM</td>
<td>New EM</td>
<td>New EM</td>
<td>New EAP</td>
<td>New EAP</td>
</tr>
<tr>
<td>L=10</td>
<td>N=200</td>
<td>-0.009 0.033</td>
<td>0.145 0.173</td>
<td>0.042 0.070</td>
<td>0.181 0.238</td>
<td>0.013 -0.024</td>
</tr>
<tr>
<td></td>
<td>N=500</td>
<td>-0.008 0.012</td>
<td>0.107 0.112</td>
<td>-0.003 0.021</td>
<td>0.129 0.143</td>
<td>0.015 -0.007</td>
</tr>
<tr>
<td></td>
<td>N=1000</td>
<td>-0.020 0.010</td>
<td>0.080 0.075</td>
<td>0.004 0.031</td>
<td>0.101 0.104</td>
<td>0.025 -0.002</td>
</tr>
<tr>
<td></td>
<td>N=2000</td>
<td>-0.011 0.008</td>
<td>0.062 0.051</td>
<td>0.000 0.031</td>
<td>0.087 0.079</td>
<td>0.004 -0.004</td>
</tr>
<tr>
<td>L=20</td>
<td>N=200</td>
<td>-0.042 -0.001</td>
<td>0.164 0.166</td>
<td>-0.017 0.056</td>
<td>0.195 0.219</td>
<td>0.055 0.005</td>
</tr>
<tr>
<td></td>
<td>N=500</td>
<td>-0.013 0.011</td>
<td>0.117 0.097</td>
<td>-0.022 0.045</td>
<td>0.157 0.144</td>
<td>0.011 -0.004</td>
</tr>
<tr>
<td></td>
<td>N=1000</td>
<td>-0.032 0.011</td>
<td>0.098 0.064</td>
<td>-0.033 0.028</td>
<td>0.135 0.097</td>
<td>0.032 -0.004</td>
</tr>
<tr>
<td></td>
<td>N=2000</td>
<td>-0.008 0.005</td>
<td>0.070 0.050</td>
<td>-0.044 0.017</td>
<td>0.117 0.067</td>
<td>-0.001 -0.003</td>
</tr>
<tr>
<td>L=40</td>
<td>N=200</td>
<td>-0.083 0.039</td>
<td>0.184 0.207</td>
<td>-0.012 0.074</td>
<td>0.193 0.227</td>
<td>0.068 -0.007</td>
</tr>
<tr>
<td></td>
<td>N=500</td>
<td>-0.048 0.036</td>
<td>0.142 0.127</td>
<td>-0.012 0.045</td>
<td>0.132 0.131</td>
<td>0.027 -0.013</td>
</tr>
<tr>
<td></td>
<td>N=1000</td>
<td>-0.049 0.038</td>
<td>0.120 0.094</td>
<td>0.001 0.042</td>
<td>0.106 0.102</td>
<td>0.044 -0.010</td>
</tr>
<tr>
<td></td>
<td>N=2000</td>
<td>-0.047 0.028</td>
<td>0.096 0.071</td>
<td>-0.009 0.030</td>
<td>0.081 0.077</td>
<td>0.032 -0.006</td>
</tr>
<tr>
<td>L=100</td>
<td>N=200 (^2)</td>
<td>-0.074 0.004</td>
<td>0.173 0.188</td>
<td>0.037 0.068</td>
<td>0.173 0.204</td>
<td>0.102 0.015</td>
</tr>
<tr>
<td></td>
<td>N=500</td>
<td>-0.068 0.010</td>
<td>0.114 0.118</td>
<td>0.009 0.038</td>
<td>0.117 0.123</td>
<td>0.078 0.006</td>
</tr>
<tr>
<td></td>
<td>N=1000</td>
<td>-0.064 0.010</td>
<td>0.105 0.085</td>
<td>0.001 0.030</td>
<td>0.088 0.089</td>
<td>0.069 0.004</td>
</tr>
<tr>
<td></td>
<td>N=2000</td>
<td>-0.047 0.012</td>
<td>0.089 0.066</td>
<td>-0.004 0.035</td>
<td>0.068 0.073</td>
<td>0.052 0.002</td>
</tr>
<tr>
<td>L=200</td>
<td>N=200</td>
<td>-0.074 0.008</td>
<td>0.188 0.215</td>
<td>0.032 0.081</td>
<td>0.177 0.214</td>
<td>0.069 -0.005</td>
</tr>
<tr>
<td></td>
<td>N=500</td>
<td>-0.048 0.012</td>
<td>0.147 0.118</td>
<td>0.021 0.043</td>
<td>0.125 0.128</td>
<td>0.049 -0.008</td>
</tr>
<tr>
<td></td>
<td>N=1000</td>
<td>-0.045 0.005</td>
<td>0.119 0.084</td>
<td>0.011 0.030</td>
<td>0.096 0.090</td>
<td>0.046 -0.002</td>
</tr>
<tr>
<td></td>
<td>N=2000</td>
<td>-0.033 0.005</td>
<td>0.090 0.064</td>
<td>0.003 0.031</td>
<td>0.072 0.070</td>
<td>0.032 -0.002</td>
</tr>
</tbody>
</table>

\(^2\) Results for the L=100 and N=200 condition were based on 26 replications instead of 30
The results for item parameters showed a more mixed pattern. The new method was negatively biased for threshold parameters, while MMLE/EM produced almost no bias. For the slope parameter, however, the new method had a smaller bias than MMLE/EM which led to a small positive bias. With regard to RMSE, MMLE/EM registered larger values when the N=200 regardless of test length but outperformed the new Bayesian update rules in other sample sizes. In the cases where MMLE/EM were preferable, the advantage was much larger for the intercept parameter while for the slope parameter the results were often very similar. Therefore, the new method was able to come close to the performance of MMLE/EM in item calibration and would even be preferred in situations with small sample sizes.

Table 6.2 presented the results from the 3 conditions of the robustness analysis. As can be seen from Table 6.2a, model misspecification had a large impact on parameter recovery. Item parameter estimates were biased for both Bayesian and MMLE/EM algorithm, and the RMSE was smaller for the new update rules. EAP showed minimal bias and smaller RMSE for θ estimates, although the RMSE was still much larger than the normal condition where the model was correctly specified. The new update rules, on the other hand, yielded positively biased θ estimates, and the bias increased with more administered items. For the two conditions where distributional assumptions were violated, according to Table 6.2b and 6.2c, the results shared a similar pattern. Overall, both estimation methods were quite robust to the varied data-generating distributions. The new update rules consistently produced larger bias and RMSE than MMLE/EM, but the results for both methods were not very different from those in the main simulation. Actually, when the slope parameter followed \( U(0.5,2) \), the ability estimates became more
### Table 6.2a: Parameter recovery of linear test when responses were generated by 3-parameter model

<table>
<thead>
<tr>
<th>Test length</th>
<th>Sample size</th>
<th>$\alpha$</th>
<th>RMSE</th>
<th>Bias</th>
<th>RMSE</th>
<th>$\beta$</th>
<th>Bias</th>
<th>RMSE</th>
<th>Bias</th>
<th>RMSE</th>
<th>$\theta$</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>L=40</td>
<td>N=500</td>
<td>0.290</td>
<td>0.444</td>
<td>0.354</td>
<td>0.466</td>
<td>-0.244</td>
<td>-0.195</td>
<td>0.278</td>
<td>0.330</td>
<td>0.115</td>
<td>-0.014</td>
<td>0.398</td>
</tr>
<tr>
<td></td>
<td>N=2000</td>
<td>0.314</td>
<td>0.429</td>
<td>0.352</td>
<td>0.435</td>
<td>-0.289</td>
<td>-0.209</td>
<td>0.298</td>
<td>0.307</td>
<td>0.071</td>
<td>-0.006</td>
<td>0.363</td>
</tr>
<tr>
<td>L=100</td>
<td>N=500</td>
<td>0.275</td>
<td>0.473</td>
<td>0.324</td>
<td>0.492</td>
<td>-0.195</td>
<td>-0.180</td>
<td>0.250</td>
<td>0.316</td>
<td>0.224</td>
<td>0.006</td>
<td>0.386</td>
</tr>
<tr>
<td></td>
<td>N=2000</td>
<td>0.324</td>
<td>0.468</td>
<td>0.341</td>
<td>0.472</td>
<td>-0.252</td>
<td>-0.180</td>
<td>0.266</td>
<td>0.288</td>
<td>0.141</td>
<td>0.001</td>
<td>0.315</td>
</tr>
</tbody>
</table>

### Table 6.2b: Parameter recovery of linear test when the slope parameter follows a uniform distribution

<table>
<thead>
<tr>
<th>Test length</th>
<th>Sample size</th>
<th>$\alpha$</th>
<th>RMSE</th>
<th>Bias</th>
<th>RMSE</th>
<th>$\beta$</th>
<th>Bias</th>
<th>RMSE</th>
<th>Bias</th>
<th>RMSE</th>
<th>$\theta$</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>L=40</td>
<td>N=500</td>
<td>-0.062</td>
<td>0.042</td>
<td>0.160</td>
<td>0.144</td>
<td>-0.033</td>
<td>0.072</td>
<td>0.164</td>
<td>0.162</td>
<td>0.024</td>
<td>-0.016</td>
<td>0.269</td>
</tr>
<tr>
<td></td>
<td>N=2000</td>
<td>-0.061</td>
<td>0.029</td>
<td>0.113</td>
<td>0.072</td>
<td>-0.046</td>
<td>0.046</td>
<td>0.111</td>
<td>0.086</td>
<td>0.029</td>
<td>-0.007</td>
<td>0.258</td>
</tr>
<tr>
<td>L=100</td>
<td>N=500</td>
<td>-0.057</td>
<td>0.014</td>
<td>0.153</td>
<td>0.130</td>
<td>-0.117</td>
<td>0.068</td>
<td>0.177</td>
<td>0.159</td>
<td>0.043</td>
<td>0.001</td>
<td>0.214</td>
</tr>
<tr>
<td></td>
<td>N=2000</td>
<td>-0.034</td>
<td>0.010</td>
<td>0.112</td>
<td>0.068</td>
<td>-0.093</td>
<td>0.058</td>
<td>0.127</td>
<td>0.095</td>
<td>0.019</td>
<td>-0.001</td>
<td>0.193</td>
</tr>
</tbody>
</table>

### Table 6.2c: Parameter recovery of linear test when ability parameter follows a skewed-normal distribution

<table>
<thead>
<tr>
<th>Test length</th>
<th>Sample size</th>
<th>$\alpha$</th>
<th>RMSE</th>
<th>Bias</th>
<th>RMSE</th>
<th>$\beta$</th>
<th>Bias</th>
<th>RMSE</th>
<th>Bias</th>
<th>RMSE</th>
<th>$\theta$</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>L=40</td>
<td>N=500</td>
<td>-0.113</td>
<td>-0.026</td>
<td>0.160</td>
<td>0.113</td>
<td>-0.024</td>
<td>0.009</td>
<td>0.163</td>
<td>0.132</td>
<td>0.037</td>
<td>-0.012</td>
<td>0.309</td>
</tr>
<tr>
<td></td>
<td>N=2000</td>
<td>-0.121</td>
<td>-0.042</td>
<td>0.137</td>
<td>0.079</td>
<td>-0.057</td>
<td>-0.019</td>
<td>0.127</td>
<td>0.084</td>
<td>0.049</td>
<td>0.002</td>
<td>0.305</td>
</tr>
<tr>
<td>L=100</td>
<td>N=500</td>
<td>-0.091</td>
<td>0.011</td>
<td>0.148</td>
<td>0.125</td>
<td>-0.010</td>
<td>0.024</td>
<td>0.132</td>
<td>0.131</td>
<td>0.063</td>
<td>-0.019</td>
<td>0.232</td>
</tr>
<tr>
<td></td>
<td>N=2000</td>
<td>-0.100</td>
<td>-0.004</td>
<td>0.125</td>
<td>0.068</td>
<td>-0.029</td>
<td>0.012</td>
<td>0.095</td>
<td>0.073</td>
<td>0.065</td>
<td>-0.006</td>
<td>0.227</td>
</tr>
</tbody>
</table>
accurate probably because the average slope/discrimination parameter was larger, meaning items were of higher quality.

To conclude, the new Bayesian update rules performed slightly worse overall than MMLE/EM in terms of estimation accuracy. The advantage of the new algorithm mainly lay in the computation speed, with more time being saved in larger datasets. It also had no convergence issues and thus allows for calibration of a large item pool with a small sample that would otherwise cause trouble with the traditional MMLE/EM approach. These benefits come at the cost of compromised measurement precision, which can still be acceptable in the low-stake assessment context. The new update rules also perform relatively well when the latent trait distribution was skewed but should be used with caution in situations where the 3-parameter model is more appropriate because in this case the new method could produce positive bias in ability estimates, especially in longer tests.

**Study 2**

The second study applied the new update rules to an online learning setting that includes multiple sessions between which changes of learner ability occur. The new algorithm was tasked with calibrating item parameters and tracking ability changes at the same time, and its performance was evaluated in combination with adaptive item selection. Last but not least, to compare the Bayesian update rules with existing procedures in psychometrics, a modified version of online calibration was also applied to the same response data as a baseline condition.

**Simulation design.**
Data generation. A total of 4 sessions were simulated with each session having 20 or 30 items, and a sample of 3,000 $\theta$s was generated from the standard normal distribution. That is, each of the 3,000 students completed 80 or 120 items in 4 different sessions with an interval in between each of them. A real-world example would be a student taking a weekly online course and taking the assignment for 4 consecutive weeks. The prior distributions and response generation methods were the same as in study 1. Within each session the response generation also followed the by-person design with the learners starting the session in random order one after another. The order effect was further negated by the fact that there were multiple sessions in which the data were entered randomly, so that a learner could be an early participant in one session and a late participant in another, thus less likely to be having a systematic advantage or disadvantage compared to others.

The item bank consisted of 500 items generated in the same manner as in Study 1. Item parameters were assumed to be constant across sessions, and all sessions shared a common item bank. The bank size was relatively small by CAT standard as with 30-item sessions the ratio of test length to bank was close to 1:4, but this is offset by the fact that in learning environment items are allowed to be delivered repeatedly for instructional purpose. It makes item bank usage more efficient and allows students to consolidate what they have learned by completing some of the items. To keep learners from memorizing the item, the same item could not appear in the same session, and an item could appear in no more than 2 sessions.

The change pattern of ability parameter was assumed to only involve between-session change and simulated as follows. The starting value was generated the same way
as study 1. Ability was assumed to be constant within each session, while after each session, a change parameter randomly sampled from $N(0.2, 0.1^2)$ was added to each ability parameter. As a result, each individual would have a different growth pattern, and a drop in ability is possible in some cases. The parameter update rules were largely the same as in Study 1 except for the following modification. Because there usually is a time interval between sessions, the uncertainty about the ability estimate would be increased during this time period. To account for the increased uncertainty of learner ability, a value was added to the posterior variance (or SD) of each ability parameter at the beginning of each new session except the first. For instance, suppose at the end of a session the posterior SD of ability of learner $i$ is updated as $\sigma_{\theta_i}$, then at the start of the next session, the posterior SD is set to be

$$\sigma_{\theta_i}' = \sigma_{\theta_i} + \tau$$  \hspace{1cm} (77)

where $\tau$ is the added SD. A similar approach was taken in previous research where the $\tau$ was either chosen a priori or estimated from the first batch of data (Glickman, 1999; Weng & Coad, 2018). In this simulation SD was modified with $\tau$ chosen to be 0.5, and separate pilot simulations found that any number between 0.25 to 0.75 would yield similar and reasonable results. Note that the added value only applied to the ability variance, while item parameter and the means of ability did not receive special adjustment when transitioning between sessions thanks to the addition of precalibrated items, which would be explained next.

**Precalibrated Items.** In study 1, the algorithm started with none of the item parameters being known. This would no longer work, however, under the setting of this
study because the changing abilities led to an issue with scaling that did not allow for simultaneous estimation of ability and item parameters. Due to the lack of a common scale, biased estimates would occur when calculating parameter estimates based on response data from multiple sessions with differing underlying latent distributions caused by the shift in population means of ability. The scale could be fixed by the use of anchor items whose parameters are precalibrated (Wang et al., 2016). The anchor items can either be a common set of linking items that appear in all sessions or completely different across sessions. In this simulation the latter is preferred because it meshes better with the adaptive nature of the online learning system. The question remains how many of the items in the bank should be precalibrated as one of the advantages of applying the new Bayesian update rules is to save the cost of item calibration and it is ideal to have as many uncalibrated items to start with as possible. Based on the recommendation that at least 20% of the items should be used for linking purpose (Kolen & Brennan, 2014), the lower bound of the proportion of precalibrated items was set to 20%. Another level at 50% was chosen as an upper bound to determine if having more known items in the bank would boost the measurement performance of the algorithm.

Although due to scaling issues the algorithm fails to achieve the original goal of doing away with any precalibrated items, if the system could still function reasonably well with as low as 20% of items being precalibrated, it would still greatly reduce the cost of item bank construction because only a small number of starting items are needed while the majority of the items can still calibrated on the fly.

**Item selection.** For the adaptive delivery of items, three item selection methods were considered. Random selection was used as a baseline condition and expected to
yield the worst results for ability estimation. Matched difficulty (MD) was used because it represented the item selection methods for ERS/Rasch models and allowed for a comparison with the 2-parameter model. It simply selected the candidate item with the smallest absolute difference between the current ability level and the item difficulty.

As is shown in Equation 18, MPWI is a Bayesian item selection criterion that integrates over the posterior distribution of learner ability. Because in this simulation the item parameters were also described by their posterior distributions, the MPWI was modified to integrate over the item parameter distributions as well

\[
MPWI = \int J_{U_j}(\theta_i, \alpha_j, \beta_j) g(\theta_i|u_{i1}, ..., u_{im}) h(\alpha_j, \beta_j|u_{i1}, ..., u_{im}) d\theta_i d\alpha_j d\beta_j
\]

where the observed information \( J_{U_j}(\theta_i) \) and posterior distribution of \( \theta_i \) \( g(\theta_i|u_{i1}, ..., u_{im}) \) are the same as in Equation 18. The posterior distribution of \( \theta_i \) in this case was \( \phi(\theta_i|\mu_{\theta_i}^{(m)}, \sigma_{\theta_i}^{(m)} \) ). On the other hand, \( h(\alpha_j, \beta_j|u_{i1}, ..., u_{im}) \) is the posterior distribution of the item parameter vector after \( m \) items and assuming independence between intercept and slope, it was found as the product of posteriors of the two item parameters: \( \phi(\alpha_j|\mu_{\alpha_j}^{(m)}, \sigma_{\alpha_j}^{(m)} \phi(\beta_j|\mu_{\beta_j}^{(m)}, \sigma_{\beta_j}^{(m)} \).

The integral in the MPWI was evaluated using Monte Carlo integration as follows: further assuming item parameters and \( \theta \) are also independent of each other,

\[
MPWI = \frac{1}{D} \sum_{d} J_{U_j}(\theta_{id}, \alpha_{jd}, \beta_{jd})
\]
where $\theta_{i_d} \alpha_{j_d}$, and $\beta_{j_d}$ are the $d$th draw from the corresponding posterior distributions with the total number of draws $D$ set to 1000.

As is discussed in Chapter 2, under 2-parameter normal ogive models, the observed and Fisher information are not exactly equivalent. However, there is evidence showing that the two types information do perform similarly in CAT with the graded response model (Choi & Swartz, 2009). In this simulation, the index was calculated based on the Fisher information defined in Equation 6 as the observed information of the normal ogive model has a very complex form and is very slow to compute. That is, $J_{U_j}(\theta_{i_d}, \alpha_{j_d}, \beta_{j_d})$ in Equation 79 was replaced by $I_{U_j}(\theta_{i_d}, \alpha_{j_d}, \beta_{j_d})$.

In previous simulation studies that involve calibrating items with discrimination parameters (Makransky & Glas, 2010; Ren et al., 2017), the calibration processes were severely affected by uneven item exposure. Some items were rarely chosen by the selection criteria and stayed in the calibration phase for an extended period of time before being put to operational use. The restrictive progressive method, an item exposure control method proposed by Wang, Chang, and Huebner (2011), was used to balance the frequencies of item delivery. Their method was modified based on the original progressive method by Revuelta and Ponsoda (1998), which added to the selection index a random component whose weight decreased as the test progressed. This made item selection at the early stages rely less on information indices that were calculated from the unreliable parameter estimates. The restrictive progressive method added a constraint which capped the exposure rate at a prespecified maximum. Combining it with the MPWI, the final selection criterion is written as
\[
\left(1 - \frac{r_j}{r_{\text{max}}} \right) \left[ \left(1 - \frac{y}{L} \right) R_j + \text{MPWI}_j^* \times \frac{b y}{L} \right]
\]

where \(y\) is the number of items already administered, and \(r_j\) is the exposure rate of item \(j\).

\(r_{\text{max}}\) is the prespecified maximal exposure rate and was set to 0.40 in this study. \(R_j\) denotes the random component and was sampled from a uniform distribution between 0 and the maximum of \(\text{MPWI}\) among all candidate items. \(b\) is a weighting coefficient that balances the trade-off between exposure control and estimation accuracy; and \(b\) larger than 1 indicates that estimation accuracy is prioritized. In this study it was set to 2, following Wang et. al. (2011).

The item selection indices were calculated for each eligible item remaining in the pool and the candidate item with the largest value was delivered as the next item. For precalibrated items, the true item parameters were used, while for the new item indices were computed from the current parameter estimates. Because the MPWI method took into account the slope/discrimination parameter, it was expected to result in better ability parameter recovery compared to the random selection or MD methods.

**Comparison with online calibration.** Because the new Bayesian method also requires the addition of precalibrated items at the beginning, the setting was comparable to that of online calibration. To see if the new algorithm could offer some improvement over the existing methods, online calibration was added as a fourth condition in addition to the three item selection methods and applied to the same data generation configurations. The operational items from online calibration were simply the precalibrated items discussed above, with their proportion also being varied at 20% and
Aspects of the online calibration design were slightly modified to accommodate some of the features of the current simulation design.

To offer a fair comparison between online calibration and the new Bayesian estimation, the datasets used for the online calibration were the same as those in the random item selection condition. That is, item delivery was also completely random for every learner. This means that the seeding locations for operational items were also random, and unlike the original online calibration design, the number of operational items for each learner was not fixed.

The major difference between the online calibration and the random selection method was how the parameters were estimated. In random selection, parameters were updated after every response, while in online calibration the parameters were estimated after each session. There was no need to obtain interim parameter estimates because no adaptive item selection or ability estimation took place in the middle of a session. After the end of a session, new item parameters were first computed by MEM as in Equation 22 with the precalibrated item parameters fixed, and then ability parameter was found by EAP using both estimated new items and the precalibrated items. To account for the changing abilities, the population means were allowed to vary across sessions and estimated during the MEM process for each session using Equation 18. Note that in the original online calibration design only operational items are used for scoring. This was not feasible in the current simulation because, in some conditions there were very few operational items (for instance, on average there were as few as 4 operational items in the 20%/20 items condition) in a session. In addition, using all items for scoring makes the online calibration estimation more comparable to the Bayesian update rules which also
score learners based on all items. The estimated new item and ability parameters were then used as the starting values of the next session, and the process was repeated until all four sessions were finished.

**Evaluation criterion.** In summary, there were a total of 2 test length (20/30 items per session) × 2 precalibrated items (20%/50%) × 4 item selection method/design (random/MD/MPWI/online calibration) =16 conditions. In each condition, the same item bank was used to generate responses in 30 different samples. The bias and RMSE of ability estimate, as well as the correlation between estimated and true abilities, would now be reported by session, while for item parameter recovery only the final estimates were considered. In addition, the posterior means and SD of ability estimates after every update were recorded for each learner, from which a trace plot showing the trajectory of mean and SD estimates during the sessions could be generated at the individual level.

Item exposure was examined by calculating the exposure rate \( r_j \) (number of administrations divided by the sample size) for every item in the bank within each sample. To determine whether items were administered evenly, the variability of item exposure was assessed by the standard deviation of the exposure rates within each replication. In addition, the frequencies, maximum, and minimum of exposure rates were also recorded for each sample, and then all these results were averaged across the 30 replications. All these measures of item exposure were reported for the entire item bank, then separately for precalibrated and new items. Because in the simulation some items were allowed to be selected twice, the average number of repeated items each learner encountered (out of the entire set of 80/120 items) was also reported to see if this repetition would lead to overexposure of some items.
Results.

Item exposure. Because the calibration precision of an item is directly affected by the number of learners to which it was administered, the results of item exposure rates are presented first. Table 6.3-6.5 showed the exposure statistics for the 3 item selection methods/designs, respectively, since the online calibration condition utilized the same data set from random selection. As expected, random selection had the most even item exposure rates with very small range and standard deviation. There was no noteworthy difference between new and precalibrated items in terms of exposure rates. With regard to repeated items, an individual learner only received an average of 9 repeated items out of the four 20-item sessions and 20 repeated items (that is, 10 unique items each repeated twice) out of the four 30-item sessions, and the proportion of repeated items was the same for new and precalibrated items.

MD resulted in more unbalanced item administration reflected by a much larger standard deviation of exposure rates. A small number of items were administered to less than 10% of learners and the minimum exposure rates were around 0.014 for the 20-item session and 0.036 for the 30-item session, which translates to a calibration sample of 42 and 108, respectively. On the other hand, the maximal exposure rate still did not exceed 0.3, showing no sign of overexposure even without exposure control measures. Unlike in random selection, some discrepancy existed between the results for new and precalibrated items. The average and minimum exposure rate were both lower for the new items, and it appeared that a disproportionate number of new items were underexposed when only 20% of items were precalibrated. The number of repeated items was more than doubled with a similar split between new and precalibrated items. Repeated items
Table 6.3: Item exposure statistics of the random selection condition

<table>
<thead>
<tr>
<th>Exposure rates</th>
<th>20-item session</th>
<th></th>
<th></th>
<th>30-item session</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RD_0.2</td>
<td>RD_0.5</td>
<td>RD_0.2</td>
<td>RD_0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>All  N  P</td>
<td>All  N  P</td>
<td>All  N  P</td>
<td>All  N  P</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 - 0.1</td>
<td>0          0 0</td>
<td>0          0 0</td>
<td>0          0 0</td>
<td>0          0 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1 - 0.2</td>
<td>500        400 100</td>
<td>500        250 250</td>
<td>0          0 0</td>
<td>0          0 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2 - 0.3</td>
<td>0          0 0</td>
<td>0          0 0</td>
<td>500        400 100</td>
<td>500        250 250</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.3 - 0.4</td>
<td>0          0 0</td>
<td>0          0 0</td>
<td>0          0 0</td>
<td>0          0 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.4-1.0</td>
<td>0          0 0</td>
<td>0          0 0</td>
<td>0          0 0</td>
<td>0          0 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>500        400 100</td>
<td>500        250 250</td>
<td>500        400 100</td>
<td>500        250 250</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>0.160       0.160 0.160</td>
<td>0.160       0.160 0.240</td>
<td>0.240       0.240 0.240</td>
<td>0.240       0.240 0.240</td>
<td></td>
<td></td>
</tr>
<tr>
<td>minimum</td>
<td>0.156       0.156 0.156</td>
<td>0.156       0.156 0.235</td>
<td>0.235       0.235 0.235</td>
<td>0.235       0.235 0.236</td>
<td></td>
<td></td>
</tr>
<tr>
<td>maximum</td>
<td>0.164       0.164 0.163</td>
<td>0.164       0.163 0.245</td>
<td>0.245       0.245 0.245</td>
<td>0.245       0.245 0.245</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>0.007       0.007 0.007</td>
<td>0.007       0.007 0.009</td>
<td>0.009       0.009 0.009</td>
<td>0.009       0.009 0.009</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repeated items</td>
<td>4.570       3.660 0.910</td>
<td>4.566       2.280 2.286</td>
<td>10.046      8.038 2.008</td>
<td>10.020      5.011 5.009</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. N=New items; P=precalibrated items, RD_0.2 means random selection method with 20% precalibrated items
Table 6.4: Item exposure statistics of the matched difficulty condition

<table>
<thead>
<tr>
<th>Exposure rates</th>
<th>20-item session</th>
<th></th>
<th>30-item session</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MD_0.2</td>
<td>MD_0.5</td>
<td>MD_0.2</td>
<td>MD_0.5</td>
</tr>
<tr>
<td>0 - 0.1</td>
<td>All</td>
<td>N</td>
<td>P</td>
<td>All</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>23</td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>0.1 - 0.2</td>
<td>451</td>
<td>358</td>
<td>93</td>
<td>456</td>
</tr>
<tr>
<td>0.2 - 0.3</td>
<td>24</td>
<td>19</td>
<td>5</td>
<td>22</td>
</tr>
<tr>
<td>0.3 - 0.4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.4-1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>500</td>
<td>400</td>
<td>100</td>
<td>500</td>
</tr>
<tr>
<td>Average</td>
<td>0.160</td>
<td>0.159</td>
<td>0.164</td>
<td>0.160</td>
</tr>
<tr>
<td>minimum</td>
<td>0.014</td>
<td>0.025</td>
<td>0.014</td>
<td>0.014</td>
</tr>
<tr>
<td>maximum</td>
<td>0.214</td>
<td>0.210</td>
<td>0.214</td>
<td>0.212</td>
</tr>
<tr>
<td>SD</td>
<td>0.036</td>
<td>0.037</td>
<td>0.031</td>
<td>0.033</td>
</tr>
</tbody>
</table>

Note. MD_0.2 means matched difficulty method with 20% precalibrated items
Table 6.5: Item exposure statistics of the MPWI condition

<table>
<thead>
<tr>
<th>Exposure rates</th>
<th>20-item session</th>
<th>30-item session</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MPWI_0.2</td>
<td>MPWI_0.5</td>
</tr>
<tr>
<td>All N P</td>
<td>All N P</td>
<td>All N P</td>
</tr>
<tr>
<td>0 - 0.1</td>
<td>147 121 26</td>
<td>157 84 73</td>
</tr>
<tr>
<td>0.1 - 0.2</td>
<td>166 145 21</td>
<td>147 87 60</td>
</tr>
<tr>
<td>0.2 - 0.3</td>
<td>167 122 45</td>
<td>176 73 103</td>
</tr>
<tr>
<td>0.3 - 0.4</td>
<td>20 12 8</td>
<td>20 6 14</td>
</tr>
<tr>
<td>0.4-1.0</td>
<td>0 0 0</td>
<td>0 0 0</td>
</tr>
<tr>
<td>Total</td>
<td>500 400 100</td>
<td>500 250 250</td>
</tr>
<tr>
<td>Average</td>
<td>0.160 0.156 0.177</td>
<td>0.160 0.152 0.168</td>
</tr>
<tr>
<td>minimum</td>
<td>0.021 0.022 0.021</td>
<td>0.020 0.021 0.020</td>
</tr>
<tr>
<td>maximum</td>
<td>0.338 0.338 0.328</td>
<td>0.337 0.304 0.337</td>
</tr>
<tr>
<td>SD</td>
<td>0.096 0.095 0.099</td>
<td>0.085 0.096 0.097</td>
</tr>
</tbody>
</table>

Note. MPWI_0.2 means MPWI method with 20% precalibrated items
made up 23 of the 80 items in the entire item set in the 20-item session condition, and 47 of the 120 total items in the 30-item session condition.

With regard to MPWI, the variability in exposure rates was increased further for both new and precalibrated items. The standard deviation was consistently larger in the 20-item condition; that is, item exposure was more unbalanced in the shorter sessions. Compared to the MD condition, many more items were underexposed with around 30% of items having exposure rates below 0.1 in the 20-item condition, and for the 30-item sessions about the same number of items had exposure rates below 0.2. The minimums, however, were marginally larger than those in the MD condition. In all conditions the maximal exposure rates exceeded 0.3, but also fell below the threshold of 0.4 specified in the exposure control procedures. The disparity between precalibrated and new items showed a pattern similar to that observed in the MD condition. Generally new items were more underexposed than the precalibrated items, but the difference was not very large and new items still received an adequately large sample for calibration. Repeated items were also comparable to the MD condition, with only a slight drop in numbers in most cases.

**Item calibration.** After all learners had finished their 4 sessions of items, estimates of intercept and slope parameters of the new items were recorded, and the recovery measures from the two session lengths are summarized in Table 6.6 and 6.7, respectively. The 30-item sessions clearly outperformed the 20-item sessions in all cases in terms of all three indices, with the largest improvements coming from the online calibration condition. In both session lengths, the slope parameters were estimated worse
Table 6.6: Parameter recovery of new items for the 20-item session condition

<table>
<thead>
<tr>
<th>Item Selection/Design</th>
<th>Precalibrated Items</th>
<th>Bias</th>
<th>RMSE</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\alpha$</td>
<td>$\beta$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>RD</td>
<td>20%</td>
<td>0.088</td>
<td>-0.119</td>
<td>0.148</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>0.079</td>
<td>-0.107</td>
<td>0.142</td>
</tr>
<tr>
<td>MD</td>
<td>20%</td>
<td>0.062</td>
<td>-0.047</td>
<td>0.231</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>0.046</td>
<td>-0.029</td>
<td>0.221</td>
</tr>
<tr>
<td>MPWI</td>
<td>20%</td>
<td>0.063</td>
<td>-0.106</td>
<td>0.219</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>0.044</td>
<td>-0.092</td>
<td>0.210</td>
</tr>
<tr>
<td>OC</td>
<td>20%</td>
<td>0.089</td>
<td>0.147</td>
<td>0.397</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>0.048</td>
<td>0.135</td>
<td>0.340</td>
</tr>
</tbody>
</table>

Note: OC=online calibration
Table 6.7: Parameter recovery of new items for the 30-item session condition

<table>
<thead>
<tr>
<th>Item Selection/Design</th>
<th>Precalibrated Items</th>
<th>Bias</th>
<th>RMSE</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>α</td>
<td>β</td>
<td>α</td>
</tr>
<tr>
<td>RD</td>
<td>20%</td>
<td>0.071</td>
<td>-0.100</td>
<td>0.121</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>0.062</td>
<td>-0.088</td>
<td>0.116</td>
</tr>
<tr>
<td>MD</td>
<td>20%</td>
<td>0.049</td>
<td>-0.035</td>
<td>0.209</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>0.033</td>
<td>-0.032</td>
<td>0.198</td>
</tr>
<tr>
<td>MPWI</td>
<td>20%</td>
<td>0.060</td>
<td>-0.070</td>
<td>0.189</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>0.043</td>
<td>-0.055</td>
<td>0.179</td>
</tr>
<tr>
<td>OC</td>
<td>20%</td>
<td>0.048</td>
<td>0.099</td>
<td>0.255</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>0.032</td>
<td>0.089</td>
<td>0.224</td>
</tr>
</tbody>
</table>
with larger bias and RMSE than the intercept parameter. This was even more obvious from the results on correlation in which all but one of the correlations for intercept were larger than 0.90, while for the slope parameters most correlations did not exceed 0.80. There was also a tendency for the Bayesian update rules to overestimate the intercept parameter and underestimate the slope parameter, while in online calibration that utilized MMLE/EM both parameters were overestimated.

In the 20-item session condition, MD had the smallest bias for the slope parameters while bias of the other three item selection/design methods were all close to or larger than 0.10 in absolute values. For intercept parameters, MD and MPWI had similar bias that were both smaller than those of random selection and online calibration. RMSE showed a different pattern in which random selection was clearly superior with all values below 0.20, followed by MPWI, whose values ranged from 0.21 to 0.24. Although having the smallest bias, MD yielded larger RMSE than random selection and MPWI, especially for the slope parameters. Online calibration performed the worst overall with all RMSE values approaching 0.40, even when the shift in ability was accounted for in the E step (the group averages were actually well recovered with bias smaller than 0.03). For correlation, random selection remained the best among all four method/designs. Online calibration had significantly worse performance on intercept parameter than the others, but for slope parameter it was MD that had the poorest results. Results from the 30-item session condition were similar. The increased session length led to smaller bias for the slope parameters with none of the absolute values exceeding 0.10. Bias of MD was still the smallest for both intercept and slope parameters while random selection had the largest values overall. For RMSE, the order was identical to the 20-item condition.
although all values were substantially smaller. Online calibration benefited the most from the increased session length and had RMSE values comparable to the MD condition. Correlation also showed some increment from the increased session length, especially for the slope parameter, although the values were still lower than 0.80 except for random selection.

The proportion of precalibrated items showed a small yet consistent effect on item parameter recovery. As expected, having more precalibrated items resulted in better recovery of the new items. This could partially be explained by the fact that more precalibrated items in the pool simply mean fewer new items needed to be estimated and thus each item would receive a larger calibration sample. However, increasing the number of precalibrated items only had a modest effect on recovery measures in many conditions, in which adding 150 precalibrated items (there were 100 precalibrated items in the bank in the 20% condition and 250 in the 50% condition) only reduced the RMSE by about 0.10. It could be said that number of precalibrated items was a relatively weak factor in deciding calibration accuracy compared to session length and item selection/design. With only 20% of items being known, the Bayesian update algorithm still functioned fairly well.

*Ability estimation.*

*Bias.* The bias of ability estimates over the four sessions is graphically displayed in Figure 6.1a and 6.1b separately for the two session lengths, with the results for the 20% precalibrated condition on the left and 50% condition on the right in each figure. The four vertical dashed lines in each figure denote the end of the sessions with respect to the total number of items administered.
Figure 6.1a: Bias of ability estimates over 4 sessions in the 20-item session condition
Figure 6.1b: Bias of ability estimates over 4 sessions in the 30-item session condition
In contrast to the linear testing scenario in Study 1, the new algorithm systematically underestimated learners’ ability across most conditions, except for the first session after which all item selection method/designs had minimal bias. This meant that at the group level the estimated ability growth was slower than the actual growth, which was set to be 0.20 on average on the $\theta$ scale after each session. It was also obvious from the figures that bias tended to accumulate as learners progressed through later sessions, except in the random condition in which the bias at the end of the last session was almost equal to the bias of the first session. As a result, random selection had the least bias overall with any of the (absolute) values smaller than 0.05. A similar pattern of increasing bias was also observed in previous simulation with a multi-session design (Park, Joo, et al., 2019). The trend was somewhat mitigated by the higher percentage of precalibrated items, in which case the bias started to level off after the second session, but not by the increased session length. Longer sessions also did not bring about much reduction in bias for the estimates from the three method/designs employing the new Bayesian update rules, while the EAP estimates from the online calibration procedures were noticeably less biased in the 30-item sessions.

Aside from random selection, online calibration performed better than the two adaptive selection methods (MD and MPWI) in the 20% precalibrated items condition, but slightly worse in the 50% condition, probably due to the higher quality of adaptation in the latter case since item parameters were generally more accurate. MPWI consistently outperformed MD regardless of session lengths and proportion of precalibrated items, and the difference actually got bigger over time.

$RMSE$. The RMSE of ability estimates was visualized in a similar fashion in Figure 6.2a and 6.2b. Like bias, RMSE in all conditions showed an overall increasing trend over time, with
Figure 6.2a: RMSE of ability estimates over the 4 sessions in the 20-item session condition.
Figure 6.2b. RMSE of ability estimates over the 4 sessions in the 30-item session condition

- 20% Precalibrated Items
- 50% Precalibrated Items
the exception of the second session where RMSE stayed equal to or became slightly smaller than the first. The trend was also flattened, although far from completely, by the higher proportion of precalibrated items, but still persisted in the face of increased session length. However, the trend was much more obvious for the estimates produced by the new Bayesian update algorithm, while the RMSE of the EAP estimates from online calibration increased at a slower pace. This was not the case in figures of bias where EAP estimates had bias that increased as fast as, if not faster than, the others. The longer session did bring down the RMSE values consistently by at least 0.05 across all conditions, while given the same session length, more precalibrated items only led to smaller values for the two adaptive item selection methods.

Random selection and online calibration showed results that were both considerably worse than their adaptive counterparts. When the session length was only 20, their results were particularly poor with all RMSE values over 0.35. With 50% precalibrated items, the two random designs had roughly equal results. But in the 20% cases, online calibration had RMSE comparable to random selection only in the first two sessions. For the last two sessions online calibration performed better since it was person to less accumulation of error. The difference between random and adaptive designs was much larger in the 50% precalibrated condition, again thanks to the higher quality item parameter estimates. In fact, in the 20% condition, online calibration and MD had almost the same results at the end of the fourth session. MD and MPWI performed almost identically for the first session, but for the remaining three sessions MPWI clearly performed the best by a margin of at least 0.02. The advantage of MPWI was actually bigger in less favorable conditions, such as shorter sessions or fewer precalibrated items. In the best case scenario of 30-item session with half of the items known, MPWI was able to register RMSE values that ranged from 0.21 to 0.25.
Figure 6.3a: Correlation between true and estimated abilities over 4 sessions in the 20-item session condition.
Figure 6.3b: Correlation between true and estimated abilities over 4 sessions in the 30-item session condition
**Correlation.** As is shown from Figure 6.3a and 6.3b, the correlations between true and estimated abilities largely followed a similar pattern to that of RMSE. MPWI performed the best among all calibration method/designs with MD being a close second, while online calibration had the worst results overall. Session length had a much larger impact on the outcome than proportion of precalibrated items. Unlike RMSE, however, correlation did not drop off at later sessions, which means that even though errors accumulated as the sessions progressed, the relative rank orders of estimated abilities remained the same. All conditions had correlation values higher than 0.92, while for MPWI and MD the correlations both exceeded 0.95.

**Trace plot.** Figure 6.4a and 6.4b shows the trajectory of the posterior means and SD, produced by the new Bayesian update algorithm, as a function of the items administered. The data came from two learners that were each chosen from the simulation sample to illustrate the process of parameter update within and between sessions. The vertical dashed lines, like previous figures, also marked the end/beginning of a session, but now the data points started from the first item, not the end of the first session. Each mean and SD estimate, which by design was updated after every item answered, is represented by a solid circle, and connected by lines to better depict the trend. The four horizontal line segments in the upper figure denoted the true ability level of the learner at the given session and were horizontal since no within-session growth was assumed.

Figure 6.4a was based on the condition of 50% precalibrated items in 20-item sessions with the items adaptively selected by MPWI. It can be seen from the red lines that the learner had a relatively high starting ability, about 1 standard deviation above the
population mean (starting ability was generated from standard normal distribution). The amount of ability growth after the first and third session was close to the average level at around 0.2, while the second session saw only below average improvement. Starting at 0.0 since no other information of the learner was known (cf. Park, Joo, et al., 2019), the ability estimates had some major fluctuations before stabilizing around the true ability level. The process moved on to the next session with the final update from the previous session serving as the starting value of the next session, and a similar pattern of fluctuating and stabilizing was observed in the remaining three sessions.

The second figure shows the trajectory of the posterior SD values of the same learner. As was predicted by Equation 64, the values were monotonically decreasing within each session. This helped explain why in the upper figure the update size of ability estimates became smaller towards the end of a session, as the update size was positively associated with the posterior variance of ability parameter, according to Equation 61. Note that in the simulation design a constant of 0.50 was added to the posterior SD at the beginning of a new session to account for the higher level of uncertainty (see Equation 77). This was reflected by the abrupt increase of SD values between sessions. Another finding worth mentioning was that the starting values in the later sessions were much lower than the first, yet the final SD value in all sessions turned out to be quite comparable.

In Figure 6.4b, the data were taken from the 30-item sessions with all other options being the same. The learner had a much lower starting ability, so the ability estimates had a much bigger fluctuations in the first session. Even though the session was
Figure 6.4a: Trace plot of the estimates and the standard deviation of a randomly chosen student’s ability under the MPWI in 20-item sessions with 50% pre-calibrated items condition
Figure 6.4b: Trace plot of the estimates and the standard deviation of a randomly chosen student’s ability under the MPWI in 30-item sessions with 50% pre-calibrated items condition.
longer, the final estimate could not come close to the true values, probably because the system could not recover from the early spike in estimates from a correct response. The next two sessions saw a smoother trajectory as the update sizes became less extreme and the ability was better estimated than the first session. A few consecutive correct responses drew the estimate away from the true values early in the fourth session, but then a string of incorrect responses towards the end brought the estimate down to correct for the early error.

The SD plot at the bottom was almost identical to the 20-item sessions where within each session SD continuously decreased. Since the session became longer, the final posterior SD was also slightly smaller at around 0.20. The 10 extra items also made it obvious to see how the curve would gradually flatten out towards the end of a session; it is reasonable to expect that if the session continued indefinitely, the SD would still continue to drop until “converging” to a lower bound.
Chapter 7: Summary and Discussion

An online Bayesian parameter estimation algorithm (Weng & Coad, 2018) was applied to an IRT-based adaptive learning environment, in which learners receive practice or quiz items that are individually matched to their changing ability level. The new update rule was originally derived based on the paired comparison models in statistics and have been recently adapted for IRT applications. It is proposed under the Bayesian framework and has the potential to simultaneously update item and ability parameters in real time, even when the ability is changing. The performance of the new algorithm was explored through two separate simulation studies. First, in order to be established as a viable alternative to the traditional IRT parameter estimation methods, the new algorithm was applied to a full response matrix from linear testing, and the resulting estimates were compared to those calculated from MMLE/EM and EAP under ideal as well as compromised conditions. The second simulation applied the algorithm to a multi-session measurement setting that mimicked a simplified online learning application and evaluated the capability of the new algorithm to dynamically monitor learner performance and calibrate item parameters. Three factors were manipulated to explore how the algorithm would perform under different setup, and an existing procedure of online calibration was added for comparison as a baseline.

The results from Study 1 provide preliminary evidence that the new update rules could be successfully implemented in educational assessment with psychometric performance comparable to the standard IRT methods. MMLE/EM and EAP still generate more accurate estimates, but the new Bayesian update rules compute much faster and had no convergence issues. The Bayesian methods were also shown to be
relatively robust to some assumption violations, such as moderately skewed ability distributions. One major caveat is that the new update rules, based on 2-parameter IRT models, probably will not work well with items that fit the 3-parameter models such as multiple-choice questions. It was shown that when there was a model misfit with data generated from 3-parameter models, the ability estimates from the new method exhibited positive bias, especially when the test is long (which means more new items are administered) and/or the sample is small. Online calibration studies, including those on continuous online calibration design, do not consider model misfit because the new items are not used for scoring purposes. Therefore, model misspecification of the new items would not cause serious implications with regard to the measurement outcomes. The same argument cannot be made for the new Bayesian update rules or the ERS-based research as the new items were used for estimating abilities while being calibrated. Therefore, it is advisable to make use of item types such as fill-in-the-blanks in adaptive learning systems based on the Bayesian update rules. If this is not possible and multiple-choice items have to be used, then there are no closed-form update equations for the pseudo-guessing parameter. It can nonetheless still be updated offline after some data have been collected after setting a starting value based on the number of response options of the item.

In study 2, the new algorithm was then applied to a multi-session adaptive learning system when the learner ability was changing between sessions. In the current setup, the update algorithm, combined with adaptive Bayesian item selection and exposure control measures, was able to achieve fast and reasonably accurate parameter estimations that were comparable to some of the existing psychometric methods.
Employing 2-parameter models improved ability and item parameter recovery over the 1-parameter model, on which the ERS was based. With exposure control measures implemented at the same time as recommended by Wauters et al. (2010), extreme item exposure distribution that occurred in some previous online calibration studies (Makransky & Glas, 2010; Ren et al., 2017) was not observed in the simulation, with only a few items being underexposed.

With regard to the three challenges of applying CAT procedures to the learning environment addressed in Chapter 3, the new Bayesian update algorithm provides some promising alternatives to the topic of item bank calibration. By calibrating item parameters during item administration, online learning programs no longer need to obtain the parameters of all items before putting them to use. In the current simulation, with only 20% of items in the bank being precalibrated, the system was still able to function properly, recovering ability parameters and achieving adaptivity on the fly. This could help save a considerable amount of time and resources in the item bank construction phase, and since the new items could in turn qualify as precalibrated items once some threshold has been met (standard error or number of administrations, for instance), only a small number of starting operational items could snowball into a sizable calibrated bank rather quickly.

One of the advantages of the new algorithm over the online calibration procedure is that only a small proportion of the item bank needs to be operational as the Bayesian update rules allow the new items to be used in scoring as well. As is mentioned before, in the 20-item session condition with 20% precalibrated items, there were on average only 4 operational items per session, in which case it would be unlikely to secure reliable ability
estimates unless the new items are used for scoring. When the online calibration procedure was modified to include new items to estimate ability parameters offline, the results turned out to be similar to the Bayesian update rules under random selection, but much worse than the new algorithm under adaptive selection.

There are continuous online calibration strategies (Fink et al., 2018; Makransky & Glas, 2010) that serve similar purposes to bypass the calibration studies while also allowing for adaptive item selection, yet these strategies are not really continuous in the sense that they specify multiple stages that are either random or adaptive. Therefore, even if at some point some of the new items already become available for adaptive item selection after being delivered to a large enough sample, they do not qualify as an operational item until the next adaptive stage is reached. In the new system, however, new items are eligible for item selection at any time except when they are excluded by external constraints like exposure controls. Some online calibration designs adopt a more flexible approach by updating the item parameters periodically after a new batch of learners (Ren et al., 2017; van der Linden & Ren, 2015). Going one step further, it would be theoretically optimal to update item parameters after every single new response and then use the updated parameter estimates in adaptive item selection, which is the case under the new Bayesian algorithm. Implementing such a design would be much more complex, however, when employing estimation methods like MEM. Note that according to Equation 21 and 22, each time updating the parameter of a new item requires the following information about all the learners that respond to this item: their responses to the new item, their responses to all the operational items they have answered, and the parameters of all those operational items. All of these values need to be fed into an
iterative algorithm like MEM for each response to a new item. In contrast, the new Bayesian algorithm is much simpler and faster because it involves a set of non-iterative equations and possesses the “memoryless” property so that only the most current response and parameter estimates are needed to compute the new estimates. The ease of computation also comes without too much loss in measurement precision.

The issue of dynamic ability tracking was also investigated in the simulations. Compared to similar research based on ERS, the new Bayesian algorithm offers a more sophisticated alternative that is more closely related to the traditional parameter estimation techniques used in IRT. In ERS, the key component of the algorithm, the update size $K$, is either fixed or calculated heuristically. The new update rules instead calculate the update size by approximating the parameter posterior means and variance using the most current response and parameter estimates, a preferable approach that is data-driven yet still can be implemented on the fly. The dynamic nature of ability change poses an extra challenge in the estimation procedure when the item parameters are also unknown. In IRT, item parameters are usually computed by the MMLE/EM algorithm which assumes a population distribution that was integrated out in the marginalized likelihood function. Oftentimes there is no issue when the abilities remain constant, but when the ability parameters are allowed to change as in the learning environment, the population distribution also shifts accordingly. This would become problematic when the same set of items is calibrated on different samples (or the same sample with changing abilities) because the underlying population distribution has already shifted and would be misspecified in the E-step for future samples unless the latent mean is also estimated along with item parameters. This is exactly what is done in the online calibration in this
simulation in which MEM with adjusted latent means was used. Although the results were still poor compared to calibration based on a single, nonchanging sample, they were nonetheless much more accurate than MEM with fixed means. Similarly, for the Bayesian update rules, simultaneous estimation of item and ability parameters while the latter are constantly changing is also problematic because the starting prior of ability parameters would become misspecified after ability changes, which in turn would lead to biased item calibration. There are two ways to rectify the issue: either to update the prior based on true ability changes or to fix the parameters of some of the items. Pilot simulations had shown that both methods would significantly improve parameter recovery, but in practice there is no way of knowing the true ability changes at the population level. Therefore, in this simulation the fixed parameter approach was adopted with varying proportions of precalibrated items, and the scaling issue was solved that makes dynamic ability tracking possible again.

Limitations

The simulation studies provided some preliminary support for the new estimation algorithm applied in an online learning system, but they still represented theoretical and oversimplified situations and had the following limitations. First, the change pattern simulated in Study 2 was stepwise and only occurred between sessions. In reality, there could be other possible change trajectories, such as linear growth within session as a function of item position or time (Eggen, 2012; Park, Joo, et al., 2019). It is unknown how the new algorithm would perform under such circumstances. Based on the ability trace plots presented in Figure 3a and 3b, at the individual level the algorithm would sometimes lag behind the true ability growth, and underestimating learner ability during
the session may compromise the adaptivity of item delivery. The actual ability gain (or loss) after an item, however, is likely to be small, and this might not be a serious issue if the estimates would eventually catch up to the true changes. Under the within-session growth scenario, a related problem is that, based on Equation 61, the update size of the posterior mean would decrease within a session because it is associated with the posterior variance that is also monotonically decreasing at the same time. This is confirmed in the trace plots where posterior means showed less abrupt changes towards the end of a session. It was mostly not problematic in the current simulation with stepwise between-session changes, but if there are, for instance, linear changes that occur within session, the algorithm would not be able to keep up with the ability growth as the session progresses. A simple fix to the issue would be to allow the posterior variance to linearly increase with time as well, as was done in Glickman (1999), and the rate of variance increase could either be inferred after data collection, or predefined like the variance adjustment at the end of the session used in this simulation. Doing so keeps the update sizes from decreasing and makes it possible to capture continuous changes in ability. It also makes the system more robust to early ability updates that go in the wrong direction, as the update size later in the session is not necessarily smaller. The downside, however, is that the ability estimates would become less stable and have more fluctuations over time.

Second, the learner ability in this simulation was assumed to be a unidimensional trait. In practice, learning materials are usually divided into multiple domains or areas. For instance, in the original paper on Math Garden (Klinkenberg et al., 2011), each item belonged to one of the four domains, which corresponded to the four algebraic operations
(addition, subtraction, multiplication, and division). The ERS used in their study, however, did not differentiate between the four domains and only modelled the general math ability. Neither did the item selection methods consider content balancing that specified the number of items in each domain that should be delivered to each student. The simulation reported here represented a simplified scenario with a single latent trait that did not involve any specific domain, but in real-life applications it is expected that the learning materials would have a more complex structure. Derivation of a multidimensional extension would be challenging as the approximations of the posterior moments may not yield closed-form equations that are convenient to compute as in the 2-parameter unidimensional model. Content balancing, on the other hand, can be easier to implement with other item selection and exposure control methods to allow for a more representative coverage of learning materials. By having additional constraints on the item selection process, it is expected that the system would perform worse in terms of adaptivity and parameter recovery.

Third, it is common practice in online calibration studies to fix the number of new items (and operational items) administered to be the same across all examinees. In this simulation, no measure was taken to impose such a constraint because new and operational items were not differentiated in the item selection or ability scoring process. As a result, fairness issues could be of concern since the quality of new items cannot be guaranteed at the beginning and some learners may receive more low-quality items than others. Item exposure data confirmed that there was indeed some variability among learners in terms of the number of precalibrated items received. For instance, in the MPWI with 20% precalibrated items and 20-item session condition, among all 30
samples simulated, the range of the number of precalibrated items each learner received was from 4 to 39, with an average of 17.65 precalibrated items out of a total of 80 and a standard deviation of 3.65. That means it is very likely that some learners receive no more than one precalibrated item in some of their sessions, which is not desirable because the parameters associated with these learners (their abilities and the new item parameters they answered) were going to be poorly estimated, especially when compared to those that received an adequate number of precalibrated items. Although fairness issues are much less likely to cause legal problems in a learning environment (Veldkamp et al., 2011), the performance of the adaptive learning system is still negatively affected. It is possible to borrow the idea of seeding locations from online calibration designs so that learners would respond to precalibrated items at designated positions of the session so that it is guaranteed that they would receive an equal number of precalibrated items. In addition, content balancing techniques could be applied to keep the number of precalibrated items the same, or at least above a lower bound, for all learners.

There are some other factors that could compromise the performance of the algorithm in the adaptive learning system. For instance, it was clear from the results of Study 2 that there is an upward trend in the bias and RMSE of ability estimates over sessions, especially when most of the items in the pool are uncalibrated. It is unknown based on the results if the bias and RMSE would continue to increase or level off after a few more sessions, but it is obvious that the accumulation of error would have a negative effect on the measurement performance in later sessions, which matters more than earlier sessions since they represented the learners’ most current abilities. It has been shown that having more precalibrated items ameliorated the upward trend to some extent. Another
potential fix to the issue is to apply more sophisticated estimation methods offline after a few sessions and reset the prior distributions accordingly.

**Future Directions**

Future studies on this topic could follow some other directions that are yet to be explored. Online learning programs differ vastly in their designs and features, and it would be interesting to see how the algorithm would perform under different specifications. Future simulation designs could be further varied in terms of factors such as the number of sessions or number of items within each session. Another possibility is to have a separate item bank for each session. In some context, the ability levels would become so different over time that the use of a single item bank across multiple sessions could not be justified. Having more than one bank would probably lead to more difficulties in item calibration since more items are being included, but it has the benefit of making the bank information better aligned with the latent distribution of the sample for each session.

In their original paper, Wendy and Coad (2018) applied the normal ogive version of GRM to analyze both simulated and real Internet ratings data, which typically have more than two response options. They proposed a separate method for estimating the threshold parameters $\gamma_c$ offline based on the empirical distribution of the responses. Therefore, the threshold parameters cannot be updated until a number of responses are collected. As is previously discussed, a similar approach could be applied to the pseudo-guessing parameter of the 3-parameter IRT model when the sessions consist of multiple-choice items. It is worth exploring how the use of other IRT models would affect the
performance of the system, given that it would be difficult to derive closed-form equations for the extra item parameters.

Item selection is an important component of both CAT and adaptive learning assessment. In the simulation studies presented above, the item selection methods that were commonly employed in CAT were directly applied without major modifications. Some other approaches that may work better in a learning environment, such as shifted-ability item selection that could potentially boost learner motivation, or KL information where a target ability level could be defined based on instructional objectives, were not considered in the simulations here. In addition, the item selection methods introduced so far have focused on measuring learner abilities with item calibration as a “by-product” in the process. If the goal of the data collection process puts a higher priority on obtaining the item parameters, then item selection should focus on maximizing the information with respect to the item parameters. In this case, findings from optimal design studies in the online calibration literature could be investigated to see if they can be applied to learning assessment when necessary.

Assessment for learning is an emerging field of research that draws much attention from academic research and the testing industry. By adapting the established procedures in psychometrics to the specific requirements of online learning, it is possible to develop new assessment models and tools that make instruction and learning in the future more intelligent, engaging, and productive.
References

https://doi.org/10.3758/s13428-018-1129-1


https://doi.org/10.1111/j.1745-3984.2001.tb01123.x


https://doi.org/10.1002/j.2333-8504.2002.tb01890.x


https://doi.org/10.1016/j.compedu.2011.02.003

https://doi.org/10.1007/978-1-4939-0317-7

https://doi.org/10.1007/s11336-016-9513-1

https://doi.org/10.1017/CBO9781107415324.004


https://doi.org/10.3389/fpsyg.2019.00620

Park, J. Y., Joo, S.-H., Cornillie, F., van der Maas, H. L. J., & Van den Noortgate, W.


