# Chapter 2

# Computerized Adaptive Testing Item Selection in Computerized Adaptive Learning Systems

**Theo J.H.M. Eggen**

**Abstract** Item selection methods traditionally developed for computerized adaptive testing (CAT) are explored for their usefulness in item-based computerized adaptive learning (CAL) systems. While in CAT Fisher information-based selection is optimal, for recovering learning populations in CAL systems item selection based on Kullback-Leibner information is an alternative.

**Keywords:** Computer-based learning, computerized adaptive testing, item selection

## Introduction

In the last few decades, many computerized learning systems have been developed. For an overview of these systems and their main characteristics, see Wauters, Desmet and Van den Noortgate (2010). In so-called intelligent tutoring systems (Brusilovsky, 1999), the learning material is presented by learning tasks or items, which are to be solved by the learner. In some of these systems, not only the content of the learning tasks but also the difficulty can be adapted to the needs of the learner. The main goal in such a computerized adaptive learning (CAL) system is to optimize the student's learning process. An example of such an item-based CAL system is Franel (Desmet, 2006), a system developed for learning Dutch and French. If in item-based learning systems feedback or hints are presented to the learner, the systems can also be considered testing systems in which the main goal of testing is to support the learning process, known as assessment for learning (William, 2011). With this, a link is made between computerized learning systems and computerized testing systems.

Computerized testing systems have many successful applications. Computerized adaptive testing (CAT) is based on the application of item response theory (IRT). (Wainer, 2000; Van der Linden & Glas, 2010). In CAT, for every test-taker a different test is administered by selecting items from an item bank tailored to the ability of the test taker as demonstrated by the responses given thus far. So, in principle, each test-taker is administered a different test whose composition is optimized for the person.

The main result is that in CAT the measurement efficiency is optimized. It has been shown several times that CAT need fewer items, only about 60%, to measure the test-taker's ability with the same precision. CAT and item-based CAL systems have several similarities: in both procedures, items are presented to persons dependent on earlier outcomes, using a computerized item selection procedure. However, the systems differ because CAT is based on psychometric models from IRT, while CAL is based on learning theory. In addition, the main goal in CAT systems is optimal measurement efficiency and in CAL systems optimal learning efficiency. Nevertheless, applying IRT and CAT in item-based CAL systems can be very useful. However, a number of problems prevent the application of a standard CAT approach in CAL systems. One important unresolved point is the item selection in such systems. In this chapter, traditional item selection procedures used in CAT will be evaluated in the context of using them in CAL systems. An alternative selection procedure, developed for better fit to the goal in CAL systems, is presented and compared to the traditional ones.

## Item Selection in CAT

The CAT systems considered in this chapter presupposes the availability of an IRT calibrated item bank. The IRT model used is the two-parameter logistic model (2PL) (Birmbaum, 1968):

$$p_i(\theta) = \mathrm{P}(X_i = 1 | \theta) = \frac{\exp(a_i(\theta - b_i))}{1 + \exp(a_i(\theta - b_i))},$$

in which a specification is given of the relation between the ability, $\theta$, of a person and the probability of correctly answering item $i$, $X_i = 1 . b_i$ is the location or difficulty parameter, and $a_i$ the discrimination parameter.

In CAT, the likelihood function is used for estimating a student's ability. Given the scores on $k$ items $x_i, i = 1,...,k$ this function is given by

$$L(\theta; x_1,...,x_k) = \prod_{i=1}^{k} p_i(\theta)^{x_i} (1 - p_i(\theta)^{(1-x_i)}$$

In this chapter, a statistically sound estimation method, the value of $\theta$ maximizing a weighted likelihood function (Warm, 1989), is used. This estimate after $k$ items is given by:

$$\hat{\theta}_k = \max_{\theta}(\sum_{i=1}^{k} \mathrm{I}_i(\theta))^{1/2} L(\theta; x_1,..., x_k)$$

In this expression, the likelihood $L(\theta; x_1,..., x_k)$ is weighted by another function of the ability, $\mathrm{I}_i(\theta)$. This function, the item information function, plays a major role in item selection in CAT. In CAT, after every administered item, a new item that best fits the estimated ability is selected from the item bank. The selection of an item is based on the Fisher information function, which is defined as $\mathrm{I}_i(\theta) = E((\partial L(\theta; x_i) / \partial\theta) / L(\theta; x_i))^2$. The item information function, a function of the ability $\theta$, expresses the contribution an item makes to the accuracy of the measurement of the student's ability. This is readily seen, when it is realized that the standard error of the ability estimate can be written in terms of the sum of the item information of all the administered items:

$$se(\hat{\theta}_k) = 1 / (\sum_{i=1}^{k} I_i(\hat{\theta}_k))^{1/2}.$$

The item with maximum information at the current ability estimate, $\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}$, is selected in CAT. Because this selection method searches for each person the items on which he or she has a success probability of 0.50, we will denote this method by FI50.

## Item Selection in CAL Systems

Item selection methods in traditional CAT aim for precisely estimating ability; in CAL systems, however, optimizing the learning process, not measuring, is the main aim. Although learning can be defined in many ways, an obvious operationalization is to consider learning effective if the student shows growth in ability. In an item-based learning system, a student starts at a certain ability level, and the goal is that at the end his or her ability level is higher. The ultimate challenge is then to have an item selection method that advances learning as much as possible.

The possible item selection method explored here is based on Kullback-Leibner (K-L) information. In K-L information-based item selection, the items that discriminate best between two ability levels are selected. Eggen (1999) showed that selecting based on K-L information is a successful alternative for Fisher information-based item selection when classification instead of ability estimation is the main testing goal.

K-L information is in fact a distance measure between two probability distributions or, in this context, the distance between the likelihood functions on two points on the ability scale. Suppose we have for a person two ability estimates at two time points $\theta_{t1}$ and $\theta_{t2}$.

Then we can formulate the hypotheses that H0: $\theta_{t1} = \theta_{t2}$ against H1: $\theta_{t1} < \theta_{t2}$. H0 means that all observations are from the same distribution, and if H1 is true, this means that there is real improvement between the two estimates in time. The K-L distance between these hypotheses is given $k$ items with response $\underline{x}_k = (x_1, ..., x_k)$ have been administered is

$$K(\hat{\theta}_{t2} \| \hat{\theta}_{t1}) \equiv E\left[\ln \frac{L(\theta_{t2}; \underline{x}_k)}{L(\hat{\theta}_{t1}; \underline{x}_k)}\right] = \sum_{i=1}^{k} E\left[\ln \frac{L(\theta_{t2}; \underline{x}_i)}{L(\hat{\theta}_{t1}; \underline{x}_i)}\right] = \sum_{i=1}^{k} K_i(\hat{\theta}_{t2} \| \hat{\theta}_{t1}).$$

If we now select items that maximize this K-L distance, we select the items that maximally contribute to the power of the test to distinguish between the two hypotheses: H0, the ability does not change, versus H1, growth in ability, or learning, has taken place.

In practice, there are several possibilities for selecting the two points between which the K-L distance is maximized. In this chapter, we will study selection using the two ability estimates based on the first and the second half of the administered items. (See, Eggen, 2011, for other possibilities.) Thus, if the number of administered items is $cl$ (the current test length), the next item selected is the one that has the largest the K-L distance at $\hat{\theta}_{t_2}(x_{cl/2}, x_{1+cl/2}, ..., x_{cl})$ and $\hat{\theta}_{t_1}(x_1, x_2, ..., x_{cl/2}\big)$. This selection method is denoted by K-Lmid.

Item selection methods based on the difficulty level of the items are often considered in CAL systems. Theories relate the difficulty of the items to the motivation of learners and possible establishing more efficient learning for students (Wauters, Desmet, & Van den Noortgate, 2012). Thus, an alternative item selection method giving items with a high or low difficulty level will be studied. If we select in CAT items with maximum Fisher information at the current ability estimate, with a good item bank items will be selected for which a person has a success probability of 0.50. Bergstrom.

Lunz and Gershon (1992) and Eggen and Verschoor (2006) developed methods for selecting easier (or harder) items while at the same time maintaining the efficiency of estimating the ability as much as possible. In this chapter, we consider selecting harder items with a success probability of 0.35 at the current ability estimate. (We will label this method FI35.)

### Comparing the Item Selection Methods for Possible Use in CAL Systems

In evaluating the usefulness of selection methods in CAL systems, simulation studies have been conducted. In these simulation studies, it is not possible to evaluate the item selection methods regarding whether the individuals' learning is optimized.

Instead, only the possibility of recovering learning is compared. If learning takes place during testing, the ability estimates should show that.

In the simulation studies reported, the item bank used consisted of 300 items following the 2PL model with $\beta\sim$ N(0,0.35) and $ln\alpha\sim N(1,0.3)$. Testing starts with one randomly selected item of intermediate difficulty and has a fixed test length of 40 items. In the simulation, samples of $j=1,....,N$ =100.000 abilities were drawn from the normal distribution. Three different populations were considered representing different learning scenarios.

1. Fixed population: all simulees are from the same population and do not change during testing: $\theta$~N (0,0.35).
2. The population shows a step in growing in ability: in the first 20 items, $\theta$~N (0,0.35); after that, from item 21 to 40 $\theta$~N ($\delta$,0.35). $\delta$>0 represents the learning step that took place.
3. The population is growing linearly: $\theta$ is drawn from the normal distribution with increasing mean with the item position $\ell$ in the test: $\theta$ ~N ($\ell.\delta$/40, 0,0.35)

To evaluate the performance of item selection methods in a CAT simulation, the root mean square error of the ability after administering $\ell$ items is commonly used:

$$rmse\theta(\ell)=(\sum_{j=1}^{N}\left(\theta_j^{\ell}-\hat{\theta}_j^{\ell}\right)^2/N)^{1/2}.$$

This is a useful criterion for evaluating the estimation accuracy of the ability when simulees with fixed abilities are considered.

However, if we want to evaluate the recovery of a growing ability, a related measure, the root mean square of the difference in abilities $rmse\,\delta(\ell)$, is more appropriate. If $\delta^{\ell}_{j} = \theta^{\ell}_{2j} - \theta^{\ell}_{1j}$ is the difference between the true abilities on two points in time and $\hat{\delta}^{\ell}_{j} = \hat{\theta}^{\ell}_{2j} - \hat{\theta}^{\ell}_{1j}$ is the difference between the estimated abilities on the two time points in time, this is given by

$$rmse\,\delta(\ell) = (\textstyle\sum_{j=1}^{N}\left(\delta^{\ell}_{j} - \delta^{\ell}_{j}\right)^{2} / N)^{1/2}.$$

## Results

The results for comparing the item selection methods in the fixed population at the full test length of 40 items is given in Table 1. This confirms what was expected. In CAT, selecting items with maximum information at the current ability is the most efficient. The difference with random item selection is huge, while we lose some efficiency when we select harder items. The item selection developed for the cases in which learning takes place hardly causes any loss in efficiency when the population is not increasing.

**Table 1** $rmse\,\theta(\ell)$ at full test length for a fixed population

| Selection | $rmse\,\theta(40)$ |
|---|---|
| FI50 | 0.0972 |
| FI35 | 0.0989 |
| KL-MID | 0.0974 |
| Random | 0.1547 |

If we consider in the fixed population the $rmse\,\theta(\ell)$ as a function of the test length, then for all selection methods this is decreasing with the test length quickly approaching the maximum accuracy to be reached (in this example, about 0.09 at about 35 items). In Figure 1, $rmse\,\theta(\ell)$ is shown in a population that is growing linearly during testing.

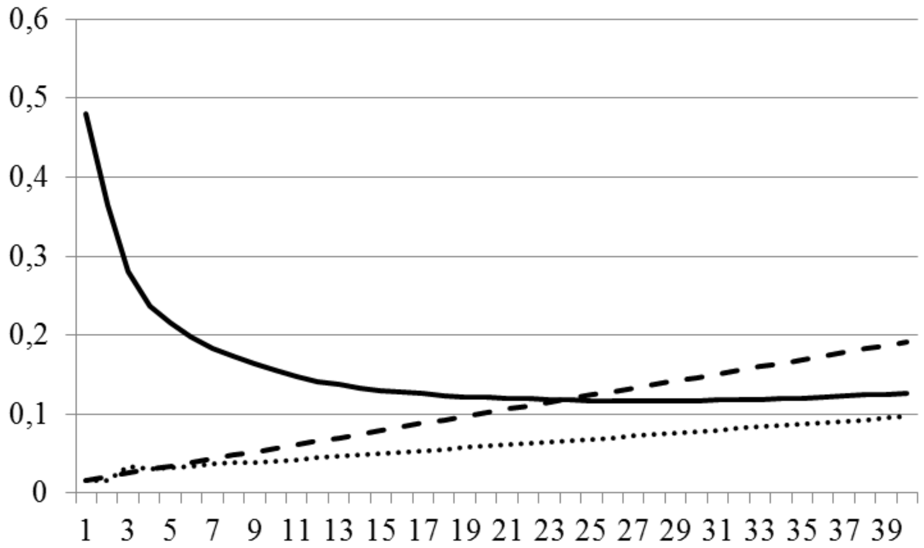

**Figure 1** True ability, estimated ability and $rmse\,\theta(\ell)$ in population growing linearly $\delta = 0.175$

In Figure 1, the dashed line (---) gives the true (growing) abilities, and the points (…) are the estimated abilities; $rmse\,\theta(\ell)$ is given by the solid line ( — ). In this situation, where the estimated abilities always lag behind the development of true ability, the $rmse\,\theta(\ell)$ is first decreasing and later increasing with the test length. This illustrates that it cannot be a good criterion for judging the recovery of growth in ability.

Therefore, the selection methods are compared on the $rmse\,\delta(\ell)$. In all the simulation studies conducted, $rmse\,\delta(\ell)$ is monotone decreasing with growing test length. Thus, the results in Table 2 are given for the full test length of 40 items. The results refer to the situation in which the increase in ability during testing is 0.175 (0.5 SD [standard deviation] of the true ability distribution).

**Table 2** $rmse\,\delta(\ell)$ for fixed, stepwise and linearly growing population

| | Growth scenario | | |
|---|---|---|---|
| Selection | Fixed | Step | Linear |
| FI50 | 0.192 | 0.196 | 0.195 |
| FI35 | 0.195 | 0.199 | 0.197 |
| KL-MID | 0.192 | 0.195 | 0.194 |
| Random | 0.303 | 0.306 | 0.304 |

To recover growth in ability, the differences between the item selection methods show about the same pattern as reported on the measurement accuracy in a fixed population: random item selection performs badly, while selecting harder items also has a negative influence on the $rmse\,\delta(\ell)$. The differences between selecting with FI50 and the KL-mid method are small; however, in populations where there is growth in ability the selection method based on the K-L information performs a bit better. Figure 2 shows where for which ability levels in the population with linear growth the small difference between the FI50 and KL-Mid method occurs.

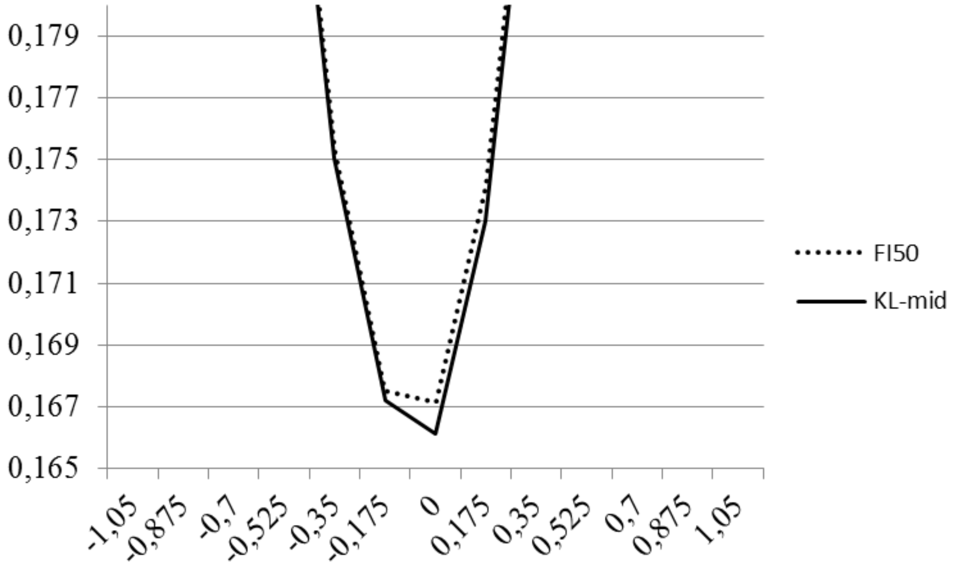

**Figure 2** $rmse\,\delta(\ell)$ for $\ell$ =40 for FI50 en KL-mid item selection as function of ability

Figure 2 shows there are only very small differences in performances for abilities around the mean of the population, which could be possibly be due to the item bank, which was constructed so that the distribution of difficulties is centered on the population mean of 0. Differences between the item selection method may appear only when there are many items of the appropriate difficulty available.

## Discussion

In computerized adaptive testing, item selection with maximum Fisher information at the ability estimate determined during testing based on the given response is most efficient for measuring individuals' abilities.

In this chapter, a K-L information-based item selection procedure was proposed for adaptively selecting items in a computerized adaptive learning system. It was explained that selecting items in this way perhaps better fits the purpose of such a system to optimize the efficiency of individuals' learning.

The proposed method was evaluated in simulation studies with the possibility of learning growth recovery as measured by the $rmse\,\delta(\ell)$ expressing the accuracy by which real growth in ability between two points in time is also estimated to be there. The results clearly showed that randomly selecting items and selecting harder items, which could be motivating in learning systems, have a negative effect. The differences between the Fisher information method and the KL information method for item selection were small.

The simulation studies reported in this chapter cover only a few of the conditions that were explored in the complete study (Eggen, 2011). In these studies, the differences were also explored

- for a very large (10.000 items) one-parameter Rasch model item bank;
- for varying populations distributions with average abilities one or two standard deviations above or below the population on which was reported here and which has a mean at the mean of the difficulty of the items in the item bank;
- for varying speed in the growth during testing (small, intermediate, large);
- for three other K-L information-based selecting methods evaluated at two different ability estimates; for instance, ability estimated based on only the first items and the estimate based on all items; and
- for different maximum test lengths.

In all these conditions, the same trends were observed. In populations that grow in ability, the K-L information selection method performs better than Fisher information-based selection methods in recovering growth. The differences however are small. In 50 repeated simulations with 10.000 students, the statistical significance of the differences was not proved.

The reasons for the lack of significant improvement are not clear. Maybe there are despite the trends only significant improvements to be expected in certain conditions not studied yet. Another reason could be that all selection methods depend on the accuracy of the ability estimates.

The K-L information-based item selection could suffer more from this than Fisher information-based selection because with K-L information two ability estimates based on only parts of the administered item sets are needed.

The first exploration of a combination of both methods consisting of using for item selection Fisher information in the beginning of the test and K-L information when at least a quarter of the total test length is administered has been conducted (Eggen, 2011). However, in this case the method performed only a tiny bit better. Nevertheless, the combination method deserves more attention.

Finally it is recommended to combine the K-L information item selection method with better estimation methods during test administration. In this context, the suggestion made by Veldkamp, Matteucci, and Eggen (2011) to improve the performance of the selection method by using collateral information about a student to get a better prediction of his or her ability level at the start of the test could be useful.

However, even more important for the practice of computerized adaptive learning system is having better continuously updated estimates of the individual's ability. The application of the dynamic ability parameter estimation approach introduced by Brinkhuis and Maris (2009) is very promising and should be considered.

**References**

Bergstrom, B.A., Lunz, M.E., & Gershon, R.C. (1992). Altering the level of difficulty in computer adaptive testing. *Applied Measurement in Education, 5,* 137-149.

Birmbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.). *Statistical theories of mental test scores* (pp 397-479). Reading, MA: Addison Wesley.

Brinkhuis, M.J.S. & Maris, G. (2009). *Dynamic parameter estimation in student monitoring systems.* Measurement and Research Department Reports (Rep. No. 2009-1). Arnhem: Cito.

Brusilovsky, P. (1999). Adaptive and intelligent technologies for Web-based education. *Künstliche Intelligenz, 13,* 19-25.

Desmet, P. (2006). L'apprentisage/enseignement des langues á l'ére du numérique: tendances récentres et défis. *Revue francaise de linguistique appliquée, 11,* 119-138.

Eggen, T.J.H.M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23,* 249-261.

Eggen, T.J.H.M. (2011, October 4). *What is the purpose of the Cat?* Presidential address Second International IACAT Conference, Pacific Grove.

Eggen, T.J.H.M. & Verschoor, A.J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement, 30,* 379-393.

Van der Linden, W.J. & Glas, C.A.W. (Eds). (2010). *Elements of adaptive testing.* New York, Springer.

Veldkamp, B.P., Matteucci, M., & Eggen, T.J.H.M. (2011). Computer adaptive testing in computer assisted learning. In: Stefan de Wannemacker, Geraldine Claerebout, and Patrick Decausmaeckers (Eds.). *Interdisciplinary approaches to adaptive learning; a look at the neighbours. Communications in Computer and Information Science, 126,* 28-39.

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer*. London: Erlbaum.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427-450.

Wauters, K., Desmet, P., & Van den Noortgate, W. (2010). Adaptive item-based learning environments based on item response theory: possibilities and challenges. *Journal of Computer Assisted Learning, 26,* 549-562.

Wauters, K., Desmet, P., & Van den Noortgate, W. (2012). Disentangling the effects of item difficulty level and person ability level on learning and motivation. Submitted to *Journal of Experimental Education.*

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation, 37,* 3-14.