Graduate
Management
Admission
Council®

Choices in CAT Models in the Context of Educational Testing

Theo Eggen
CITO and The University of Twente
The Netherlands

Presented at the CAT Models and Monitoring Paper Session, June 7, 2007

2007 GMAC® Conference on Computerized Adaptive Testing

Abstract

Procedures for the calibration of item banks used in CATs and item selection procedures in CAT algorithms meeting practical constraints are discussed. The procedures are implemented in a number of educational testing programs at Cito, the National Institute for Educational Measurement in the Netherlands. The main focus of this paper is on the measurement or psychometric aspects of the algorithms.

Acknowledgment

Presentation of this paper at the 2007 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®. A version of this paper will appear as Chapter 10: Adaptive testing and item banking, pages 199-217, in J. Hartig, E. Klieme & D. Leutner (Eds.). *Assessments of competences in educational contexts,* to be published by Hogrefe & Huber Publishers in the fall of 2008

Copyright © 2007 by the Author

Citation

Eggen, T. J. H. M. (2007). Choices in CAT models in the context of educational testing.
In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.*
Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Theo Eggen, Cito, Nieuwe Oeverstraat 50, P.O. Box 1034,
6801 MG Arnhem, The Netherlands. Email: theo.eggen@cito.nl

Choices in CAT Models in the Context of Educational Testing

Adaptive testing involves tailoring a test during testing to the trait (ability or proficiency) level of an individual examinee. In the eighties of the last century, adaptive testing evolved from a theoretical idea to a sophisticated operational form of testing in practice as the result of simultaneous developments in computer technology and in educational and psychological measurement. In computerized testing, richer item formats became available, which made it possible to build item banks that are more valid for the competencies to be measured. Furthermore, the increasing power of computers not only made it possible to present graphics and running video in items, but also non-trivial computations during testing for the selection of items and the estimation of examinee trait levels could be performed in a very short, not noticeable, period of time. On the other hand, developments in measurement theory, in particular item response theory (IRT) (Van der Linden & Hambleton, 1996), have made available the tools for computerized adaptive testing (CAT).

The main focus of this paper is the measurement or psychometric aspects of CAT, also considering practical application. This paper starts with a general description of CAT, followed by a discussion on item banking. Next, a presentation follows of the different components of a CAT algorithm, which is a set of rules that determines the way a CAT is started, continued and terminated. The relation between practical goals of testing and elements of the algorithms will be discussed. The paper concludes with a number of research ideas.

Basic Elements of Adaptive Testing

In CAT, the construction and administration of the test is computerized and individualized. For every examinee, a different test is constructed by selecting items from an item bank tailored to the trait level of the examinee as demonstrated by the responses given thus far during testing. The primary motive for CAT is efficiency. Compared to paper-based testing, there is, of course, an increase in the efficiency of testing due to computerizing the testing procedure, but emphasis is on the gain in measurement efficiency. It has been shown that CATs need fewer items to measure with the same precision. Compared to a linear, non-adaptive, test only about 50% to 60% of the number of items are needed. Since the publication of the basic ideas on modern CAT by Lord (1970), the educational and psychometric community has produced numerous articles and books on this subject. Recently, a number of books has presented overviews of the literature. Wainer (2000) gives the historical development and the basics of CAT and describes the possibilities for building, maintaining, and using CATs. Van der Linden & Glas (2000) is a compilation of recent psychometric research on CATs. Finally, Parshall, Spray, Kalohn & Davey (2002) gives a more practical overview of issues in computerized testing and CAT.

Differences Between CATs and Linear Computerized Tests

In a linear computer-based test (CBT) the testing is computerized, and before the test administration the order of the presentation of the items in the test is fixed. The main distinctive features of CATs compared to linear CBTs are:

1. Every individual gets a potentially unique test; that is, the content as well as the length of the test can differ from person to person.
2. A CAT is optimized for the individual, which has at least two favorable consequences:

  a. Measurement efficiency is enhanced, since fewer items are needed to achieve the same precision;
  b. Every examinee can be challenged at his of her own level, which has generally a stimulating effect and is experienced as pleasant.
3. Fewer items are needed to measure each individual, so individual examinees, test constructors, and test organizers can save time and/or money.
4. Since items and tests are developed within the framework of a sound test theory (IRT), the test has a number of known characteristics and therefore probably better measurement quality.

In addition to these distinctive advantages, CAT has several advantages in common with linear CBTs. Compared to paper-based testing, the most important advantages are:

1. Direct scoring and feedback is possible and the scoring is objective (errorless).
2. New item types become available; that is, not only new formats of answering items, but richer, more authentic, items can be presented.
3. Test administration is more efficient.
4. Possibilities to be flexible in test scheduling and location;
5. Better test security.
6. Examinees are more motivated.
7. More possibilities to apply modern test theory.

Item Banking and Item Response Theory

CATs presuppose the availability of an item bank. An item bank is a collection of items that is constructed to measure a well-defined area of knowledge or expertise. In addition to the items themselves, the item bank contains various characteristics of each item. These item characteristics might relate to content or administrative information, as well as psychometric characteristics—the item parameters. The item parameters are derived through the application of IRT. In IRT a relation is specified between the non-observable trait, θ , that is measured, and the observable score on the test (items). In the case of dichotomously scored items, the most popular item response models are of the logistic type. In these models, a specification is given of the relation between the θ of a person and the probability of correctly answering item i , $X_i = 1$. The exact relationship is determined by the parameters of the items. A commonly used IRT model is the two-parameter logistic model (2PL):

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

in which b_i is the location or difficulty parameter and a_i is the discrimination parameter. A special case of the 2PL is the Rasch model (one-parameter logistic model; 1L) in which only a difficulty parameter b_i is present and all a_i are equal to the same constant. Another popular IRT model that is often applied in CAT is the three-parameter logistic model (3PL), which is an extension of the 2PL with a guessing parameter

Working with an IRT model starts with item calibration. Item calibration involves fitting an IRT model and estimating the item parameters. If the item parameters are estimated, they can be stored in the item bank and used for CAT. In a CAT, the likelihood function of an examinee's θ, plays a central role in the inference on the examinee, for example, in estimating his or her θ. Given the scores on k items $x_i, i = 1,...,k$ this function is given by

$$L(\theta; x_1,...,x_k) = \prod_{i=1}^{k} p_i(\theta)^{x_i}(1 - p_i(\theta)^{1-x_i}), \tag{2}$$

which states that the probability of getting the observed scores on the items is a function of the θ. The use of the likelihood function in CAT will be explained in more detail in the next section.

The properties of IRT that make it excellently suited to apply in CATs are the following.

1. With any subset of the items in a calibrated item bank it is possible to estimate the θ of an examinee on the same scale. It is, therefore, not necessary to administer the same items to examinees in order to obtain comparable estimates of their θ.
2. The difficulties of the items are expressed on the same scale as the θ of the examinees. It is therefore possible to adapt a test to a level of an examinee.
3. The information in an item is a function of θ and, due to that, the information function can serve as a basis for tailored item selection.

These favorable properties are, however, only true if the items considered fit the chosen IRT model.

Different Forms of CAT and Limitations

CATs can be designed in several ways. In this paper we will only consider CATs that are item-based, which means that the selection of the next item to administer is made after an item has been answered. This is sometimes called fullly adaptive testing, and can be considered as the limiting case of two-stage or multi-stage testing, which was one of the earliest (Lord, 1971) applications of (non-computerized) adaptive testing. In two-stage testing, all examinees are administered the same set of items or routing test and, on the basis of the results of this routing test, low scorers are assigned to an easier second stage (measuring) test and high scorers are administered a more difficult test. Eggen & Straetmans (2000) have shown that the gain in measurement efficiency of an item-based CAT compared to a two-stage test is considerable. Item-based CATs required 50% fewer items required then two-stage tests. Nevertheless, the construction of multistage tests with more stages and more different tests in each stage is getting more attention recently.

Another form of non-item-based adaptive testing was introduced by Wainer & Kiely (1987). They proposed to use testlets instead of items as the basis for selection and administering in a CAT. A testlet is a short test of a few items. The testlet approach was introduced as a method in which non-psychometric considerations can be taken in account in adaptive testing.

Item-based CATs as discussed in this paper can make use of the exponential growth in possibilities and performance in information and communication technology. But this does not imply that there are no limitations for successfully applying CAT. Most restraints apply to computer-based testing in general, but there are aspects of CATs that put extra demands on the available computer capacities. These demands are that during testing a complete item bank should be directly accessible and that computational procedures for θ estimation and item selection must be implemented in real time.

More general limitations in computerized testing are imposed on the item format or response format. Although there are ample possibilities for the type of items or stimuli that can be presented to candidates, there are still considerable limitations in the permissible responses of the candidates. This is due to the fact that these responses need automatic scoring. Although there are several promising developments (Clauser, 2002), routine application of automatic scoring in CBTs or CATs is not common. That is the reason why in computerized tests the items often are of a (multiple) choice or matching type, and in case the examinee can give his or her own response the allowed formats are limited—for example, a few words or the result of a computation. Another limitation of CATs related to the admissible response format, is that most CAT algorithms assume that the responses to items are dichotomously scored, correct or incorrect. Although some research results are available for adaptive testing with polytomously scored items (Dodd, De Ayala & Koch, 1995) and Van Rijn, Eggen, Hemker & Sanders, 2002), practical applications have been limited. The main reason is not the availability of suitable IRT test models, but again the practical difficulty of scoring polytomous items automatically. For this reason, CAT, in this paper, is limited to items that are scored dichotomously.

Item Banking

In order to be able to administer CATs, an item bank with a sufficient number of good items fitting an IRT model is necessary. All items should be good operationalizations of the trait that is to be measured. A good item bank, covering all relevant aspects of the trait to be measured is crucial for using a CAT. In principle, the criteria for good items do not differ from the criteria for the items in a paper-based test.

Having an item bank of high quality is necessary but not sufficient for a CAT. The items must also be calibrated under a chosen IRT model. On the basis of empirical data from administered items, an IRT model is fitted and the item parameters and parameters of one or more θ distributions in the population are estimated. Some important considerations in item calibration will be discussed next.

General Considerations on Choice of IRT Model and Estimation Method

An operational CAT assumes that for item selection and estimation of θ, the item parameters are known. In order to assume this the fit to the chosen IRT model should be good and the accuracy of the estimated parameters must be very high. In a CAT this is so important, because in item selection computations use the parameters as if they were known. Furthermore, examinees get different items and inaccuracies in item parameters could work differently for different examinees.

For achieving a high quality item calibration, the choice of a specific IRT model, the choice for an estimation method of the parameters, the model fit statistics, and the nature and the size of the examinee sample play important interrelated roles. (For detailed considerations about this, see for example, Fischer & Molenaar, 1995 and Van der Linden & Hambleton, 1996).

With respect to the choice of an IRT model, the simpler the IRT model, smaller sample sizes are needed and better statistical methods for estimating and testing model fit are available. For estimating the parameters with a reasonable accuracy and for testing the model with some power in the 1PL-, the 2PL-, and the 3PL-models, one needs respectively at least 200, 500 and 1,000 examinee answers per item. On the other hand, it is more difficult to obtain a good fit for a simpler model, which could mean that some items have to be deleted from the bank. This could threaten the validity of the item bank.

Two general likelihood-based methods are available for the estimation of the item parameters. First, there is the generally applicable marginal maximum likelihood method (MML). If we use this method, we have to assume that the sample we have is a random one from a specified distribution of θ in the population (typically, the normal distribution is assumed.) With MML, the item parameters and the parameters of the θ distribution are estimated simultaneously. This is not the case if we use conditional maximum likelihood (CML) for the estimation of the item parameters. With CML, we do not need assumptions on the θ distribution of theexaminees: one only needs samples from the population. In practice this could be very favorable, because in education real random samples are not easily obtained. Having samples drawn in more steps and in clusters does not invalidate the CML estimation method of the item parameters. However, the CML estimation method is not applicable in every model. It can be used in the 1PL-model and in a somewhat restricted form of the 2PL-model (Eggen, 1990; Verhelst & Glas, 1995). This model is implemented in the OPLM computer program (Verhelst, Glas, Verstralen, 1995), which contains also item fit statistics with proven good statistical properties. If one uses CML for estimating item parameters and model-fit testing, the calibration is to be completed by separately estimating the (parameters of a) θ distribution.

Incomplete Calibration Designs

The size of item banks in use for CATs is normally such that it is not feasible to use complete testing designs in the calibration of the bank. Therefore, only a portion of the examinees in a calibration study will be able to answer only a subset of the items. Sometimes there are just practical reasons for using incomplete designs; often efficiency is also a motivating factor for incomplete designs. Efficiency in item calibration, giving smaller standard errors of the item parameter estimates, is gained when (a priori) knowledge about the difficulty of the items and the trait levels of the examinees is used in allocating examinees to subsets of items. The same principle that is used in CAT for estimating the θ of the examinee is then used for efficiently estimating the item parameters.

For the calibration of an item bank, two basic approaches are available. The first is that items are calibrated and fitted in sections of the data collection that are complete in the sense that a group of examinees answers all the items in these sections and then the different sections are linked to the same scale by equating procedures. The second, and better, approach is to calibrate the items of all sections of the data collection at once to the same scale. The theory of item calibration in incomplete designs and available software make this second approach practically feasible.

Calibration designs should meet some general requirements. The most important is that the designs should be linked. The designs can be linked only if common items are administered to the same examinees or if common persons have been administered the same items. If this is not the case, the items cannot be calibrated to the same scale.

With respect to calibration in incomplete design, Eggen (2004) has published two relevant studies. In the first study (Eggen, 2004, p 61–96) the efficiency of CML and MML estimation in different incomplete design was compared. It was shown that CML and MML performed almost equally well in all studied designs. In the designs studied, different test booklets were always randomly assigned to the examinees in the sample, but the overlap between the test booklets was established in different ways. Methods that are often used in practice, such as balanced block designs and item interlaced designs, were studied. The results from this study give another good reason for using CML in item calibration if it is possible.

The other study (Eggen, 2004, p 98-134) considered item calibration in incomplete designs as missing data problems. By taking into account the stochastic nature of the missing data in design types that are often used in practice, the justifiability of using MML and CML procedures is given. This is important because software for item calibration can handle incomplete designs, but it does not recognize the stochastic nature of the missing data, which could lead to biased results. Three commonly used design types are used.

1. *Random incomplete design*. In random incomplete designs, the researcher decides which test form is taken by which examinees without using any a priori knowledge on the θ distribution of the examinees. Every examinee has an a priori known chance of taking one of the available test forms. In these designs, the test forms are often assembled from the item bank in such a way that the forms have an equal number of items and are parallel in content and difficulty. A test form can be randomly assigned to an examinee so that every examinee has an equal chance of getting a particular test form. Or more generally, an examinee gets a test form with a known probability such that the sum of the probabilities for all different test forms equals 1.0.
2. *Multistage testing design*[1]. In multistage testing designs, the assignment of examinees to subsets of items from the total item bank in a testing stage is based on the observed responses in the former stage. An example is that all examinees in the sample take the first stage test that is of medium difficulty. This (part of the) test is called the routing test. Examinees with high scores on the routing test are administered a more difficult subset of items from the item bank in the next stage and examinees with low scores on the routing test a more easy subset. The same procedure could be continued in more testing stages.
3. *Targeted testing design.* In targeted testing designs, the structure of the design is determined a priori on the basis of background information on the examinees. This background variable is usually positively related to the θs of the examinees. Examinees expected to have lower abilities are administered easier test forms, and examinees expected to have higher abilities are administered the more difficult forms. As in multistage testing designs, gains in precision of the estimates are to be expected. An example of a variable often used in these designs is the grade level of a student.

[1] Multistage testing is not only used for item calibration, but was in fact introduced (Lord,1971) for efficiently estimating the trait levels of examinees and can, in that case, be considered as a form of adaptive testing (see first section).

For these three designs some of the results of Eggen (2004) are summarized in the following table:

Design	CML	MML
Random	Correct	Correct
Multistage	Biased	Correct
Targeted	Correct	Biased

The table indicates whether MML and CML estimation give correct or biased results in the standard application of these estimation methods. By the standard application is meant the way CML and MML are implemented in the default options of computer programs. Although special measures are possible to obtain correct results in the situations, which are indicated as biased, correct solutions for this, however, are not always available in standard IRT software.

Practical Steps in Item Banking

Because the computer is used to administer the items in a CAT, it is strongly recommended that the same medium should be used to collect the item responses used for the calibration of the item bank. The functioning of the items is very often dependent on the medium by which they are presented. In many studies, differences between computer-based and paper-based item administrations have been shown (Vrabel, 2004). Only in cases where the item bank consists of items that are virtually the same as they are on paper, should paper-and-pencil based item administrations be considered for the initial calibration of the item bank.

The major steps in performing the calibration of an item bank are

1. Setting up the calibration design: drawing the sample, the composition of the test booklets, the assignment of booklets to the examinees in the sample.
2. Administering the items.
3. Conducting basic (classical) test and item analyses.
4. Checking IRT model assumptions, such as unidimensionality.
5. Selecting the IRT model.
6. Item parameter estimation and testing model fit, including examination of differential item functioning for relevant background variables.
7. Initial establishment of the item parameters in the item bank that will be used for the CAT.

If a CAT is in use, the data of the real CAT administrations can be used to update the item parameter estimates and to reconsider model fit. This is especially needed when the initial item calibration was based on paper-and-pencil administrations and also in the case in which the quality of the sample used in the calibration was less than optimal. By using the results of CAT administrations, the parameter values of the items are better established and can be tracked and maintained. In this way, for example, one can detect that items have to be deleted from the bank as the result of changes in a curriculum or simply because the items become known to the examinees.

Having the facility of data return to the item bank after a CAT administration also opens the opportunity to add new items to the item bank. Because items become known, they might have to be replaced from time to time. Instead of organizing separate item calibration administrations for this, one could also use so called seeding designs. The idea of seeding is that somewhere during a session in which a CAT is administered, a few new items also are administered. The data on these items are not used for estimating examinee θs, but only for the estimation of the parameters of the new items. Although the design for the administration of these new items has to be constructed with care and standard likelihood-based estimation methods have to be replaced by Bayesian methods (Van der Linden & Hambleton, 1996), this is an ideal procedure for maintaining the psychometric quality of the item bank.

Algorithms in CATs

CATs are governed by a testing algorithm. The algorithm is a set of rules that determine the way CATs are started, continued, and terminated. Figure 1 shows a schematic representation of a CAT algorithm.

The choices for certain elements in CAT algorithms are also determined by the purpose of the test. Educational testing can have many practical purposes. From the perspective of the test users a division in selection, placement, certification, licensing, monitoring of progress in proficiency and diagnostic testing can be made. From the measurement point of view it generally suffices to distinguish two main aims of testing:

1. *Estimation:* The goal is to get an estimate of the competence, the ability, or proficiency of a person on a well-defined domain on a one-dimensional scale. Traditionally, CATs are designed to achieve this goal as quickly and/or as precisely as possible.
2. *Classification:* The goal is to determine in which of a limited number of competency or proficiency classes a person belongs. In this case, on the θ scale one or more border or cutting points are given that decide to which category a person belongs. A precise estimate of the person’s θ is not important, but rather the correct classification into a category.

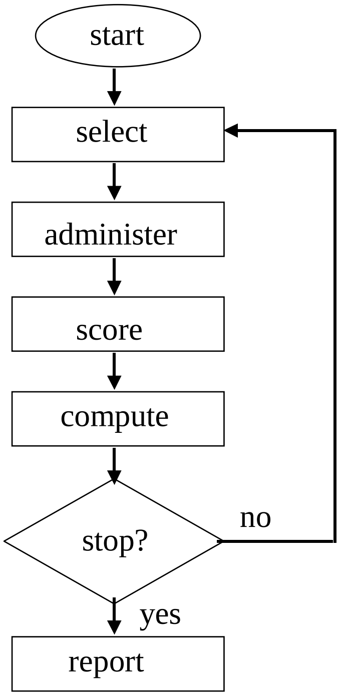
In the selection and computation part of the CAT algorithms especially, the purpose of testing can be of major influence for the choice of an optimal algorithm.

Start

In this phase, the starting item(s) in the CAT is (are) selected. As there is in general no information available on the trait level of the examinee, one could start with:

1. One or more randomly selected items from the item bank
2. One or more randomly selected items from certain sub-domain of the item bank. One could decide to start with items that have a certain common instruction for the examinees. An example is to start in a test of mathematics with a few items in which paper-and-pencil is not permitted to make computations.

Figure 1. Schematic Representation of an Adaptive Test



3. One or more randomly selected items from a certain subset of the item bank, based on psychometric characteristics of the items. An example is that one starts with a random selection of a few easy items.

If there is information available on the trait levels of the examinee before testing, then selecting the first items on the basis of that information is, of course, possible.

Select

After the administration of every item, an item selection procedure is implemented. From the item bank, an item is chosen that is in accordance with the answers given by the examinee thus far: the test is adapted or tailored to the trait level of the tested examinee. The item selection procedure is responsible for the gain in efficiency that can be reached with a CAT. It is based on an information concept of which the basic idea is that the item, which promises to give the most information on the examinee‘s θ as demonstrated thus far in the test is administered next. The criteria for item selection will be described in more detail later.

Administer and Score

These phases in running the CAT algorithm will be clear: the item is presented, answered and scored.

Compute

In the computation phase, the scores of the examinee are processed. Statistical procedures, discussed in more detail in the next section, determine on the basis of the scores on the items the estimated θ and an indication of its accuracy.

Stop

After the administration of each item, a decision is made whether a new item is to be selected or if testing can be stopped. Criteria for stopping are:

1. The accuracy of the estimate of the θ of the examinee;
2. The accuracy of the decision for classifying the examinee;
3. The maximum (and possibly also the minimum) practically available testing time or number of items that can be administered.

The last criterion, in combination with one of the first two criteria, is often chosen.

Report

Although reporting is sometimes postponed, a CAT frequently ends with a report on the results on the CAT to the examinee. Possible reports are:

1. A Report of the estimated θ, with its standard error, on a selected scale allowing for a norm-oriented interpretation;
2. A report of the estimated θ, with its standard error, on a selected scale allowing for a domain-oriented interpretation;
3. A simple (graphical) report on the category in which the examinee is classified.

Computation in the CAT Algorithm

In this phase, the θ of the examinee is estimated on the basis of the answers to the items. As was already mentioned, the algorithms make use of the likelihood function of θ, $L(\theta; x_1,..., x_k)$ .This likelihood function can be used in two different statistical procedures. The first is statistical estimation, which can be applied in any CAT application. The other is statistical testing which is a good alternative in case the goal of testing is the classification of a examinee in one of a few categories.

Estimation

The goal in many CATs is to estimate θ with as few items as possible and with predetermined accuracy. An estimate that is often used is the value of θ that maximizes the likelihood function, $L(\theta; x_1,..., x_k)$. A statistically better estimation method was introduced by Warm (1989). This estimation method maximizes a weighted likelihood function. The Warm θ estimate after k items is given by:

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

In this expression, the likelihood $L(\theta;x_1,...,x_k)$ is weighted by another function of θ, $\mathrm{I}_i(\theta)$. This function, the item information function, plays a major role in the selection of the items (see next section).

After every item, not only the estimate of θ but also the standard error of the estimate becomes available, $\mathrm{SE}\left(\hat{\theta}_k\right)$, which is an indication of the accuracy by which θ is estimated. In CATs, $\mathrm{SE}\left(\hat{\theta}_k\right)$ is often used in a stopping criterion, that is, if $\mathrm{SE}\left(\hat{\theta}_k\right)$ is below a certain level, testing is stopped.

If estimation is used in a classification problem, the θ estimate and its standard error are used for constructing confidence intervals for the true θ of the examinee:

$$\left[\hat{\theta}_k - \gamma \cdot \mathrm{SE}(\hat{\theta}_k), \hat{\theta}_k + \gamma \cdot \mathrm{SE}(\hat{\theta}_k)\right]. \tag{2}$$

The constant γ has to be specified. Depending on the level of the confidence interval, in classification CATs testing is stopped as soon as no cutting point of the classification is in the confidence interval.

Statistical Testing

Statistical testing in CAT algorithms can be applied only when the testing goal is the classification in one of a few categories. Consider the case where we have only one cutting point, θ_0, on the latent trait scale. For example, a θ below the cutting point fails an exam and above the cutting point the examinee passes the exam. A small region, with width δ, on both sides of this point is the so-called indifference zone. The indifference interval expresses the fact that, due to measurement errors, making the correct decision about examinees very near the cutting point can never be guaranteed. The problem is then formulated as the statistical test of the following hypotheses:

$$\mathrm{H}_0 : \theta \le \theta_0 - \delta \text{ against } \mathrm{H}_1 : \theta \ge \theta_0 + \delta. \tag{3}$$

Next, the acceptable decision error rates are specified:

P(accept H_0 | H_0 is true) $\ge 1-\alpha$ and P(accept H_0 | H_1 is true) $\le \beta$, in which α and β are small constants. The test meeting these decision error rates uses as the test statistic computed as the ratio of the likelihood function under H_1 and H_0:

$$\mathrm{LR}_k(\theta_2;\theta_1) = \frac{\mathrm{L}(\theta_2;x_1,...,x_k)}{\mathrm{L}(\theta_1;x_1,...,x_k)} \tag{4}$$

and has the following procedure:

if $\beta/(1-\alpha) < LR_k(\theta_2;\theta_1) < (1-\beta)/\alpha$ administer another item;

if $\mathrm{LR}_k(\theta_2;\theta_1) \le \beta/(1-\alpha)$ accept H_0;

if $\mathrm{LR}_k(\theta_2;\theta_1) \ge (1-\beta)/\alpha$ reject H_0.

The null hypothesis, H_0 is rejected with large values of the likelihood ratio. In that case the answers are far more likely under H_1 then under H_0. H_0 is accepted according to a similar, but opposite, reasoning.

The procedure described above, is readily generalized to the case in which there are three instead of two categories. For the generalization and more details of the statistical testing procedure see Eggen (2004).

Item Selection in the CAT Algorithm

In a CAT, after every administered item, a new item is to be selected from the item bank. The item being selected is the item that best fits to the θ level of the examinee as estimated at that point in the test. As a rule, the selection of an item is based on the Fisher information criterion. This method is optimal for CATs in which the main goal is θ estimation. This method also functions well in classification problems and is the preferred method in practice. Formally, the Fisher item information function is defined as

$$\mathrm{I}_i(\theta) = \mathrm{E}\left(\frac{\partial L(\theta; x_i) / \partial \theta}{L(\theta; x_i)} \right)^2 , \tag{5}$$

which is the (statistical) expectation of the squared relative change of the likelihood function of θ. The item information function, a function of θ, expresses the contribution an item can give to the accuracy of the measurement of an examinee. This is readily seen, when it is realized that the standard error of the θ estimate can be written in terms of the sum of the item information of all the administered items:

$$\mathrm{SE}(\hat{\theta}_k) = \frac{1}{\sqrt{\sum_{i=1}^{k} I_i(\hat{\theta}_k)}} . \tag{6}$$

The larger the item information, the smaller is the contribution to the standard error. With dichotomous items, the information function is a function of θ with only one maximum. For example, in the 2PL (and in the 1PL with $a_i = 1$) the function is given by

$$\mathrm{I}_i(\theta) = a_i^2 p_i(\theta)\left[1 - p_i(\theta)\right] = \frac{a_i^2 \exp\left[a_i(\theta - \beta_i)\right]}{\left\{1 + \exp\left[a_i(\theta - \beta_i)\right]\right\}^2} \tag{7}$$

It can be seen that for each item in these IRT models, the information function reaches the maximum at the value of the difficulty parameter of the item, $\boldsymbol{\theta} = \boldsymbol{\beta}_i$. It can also be seen that the discrimination parameter, a_i, has a strong influence on the value of information.

In item selection, the item with maximum information at the current θ estimate is selected. This means that after $\hat{\boldsymbol{\theta}}_k$ has been determined, the information value for every not yet administered item in the item bank is computed at this estimate value and the item with the maximum information value is administered next in the CAT.

In the CAT literature, several alternative item selection procedures have been proposed. If one favors Bayesian estimation procedures, a Bayesian selection procedure is appropriate (Van der Linden, 1998). Furthermore, it has been shown that for classification problems with one cutting point, selecting items with maximum Fisher information at that point gives somewhat better results then selecting them at the current θ estimate. The same is true of classification problems in general, where selecting on the basis of Kullback-Leibler information (Eggen, 1999) generally gives somewhat better results. However, these alternative methods have as a drawback that the test is not constantly adapted to the examinee's θ: there is only one (or a few more in the case of more cutting points in the classification) fixed order of preference of administering the items. Besides of being optimal in CAT with estimation, this is another reason why maximum Fisher information selection at the current θ estimate is the most popular item selection method in practical application.

Item Selection and Item Difficulty

The most frequently mentioned advantages of CAT are the gain in measurement efficiency and the fact that each examinee is challenged at his or her own trait level because items that are too difficult or too easy for the examinee will never be administered. But what is the difficulty of these items? It was already noted that in the 1PL and the 2PL models the information function reaches its maximum at the value of the difficulty parameter of the item, $\boldsymbol{\theta} = \boldsymbol{\beta}_i$. It is easily checked that at this value of θ the probability of a correct answer on the item is 0.50. So if one selects items that have maximum information at the θ estimate, items will be chosen for which the examinee has a probability of about 0.50 of answering the item correctly. Thus, as a rule, examinees taking a CAT will answer about half of the items correctly. Although the difficulty of the items is taken into account in the scoring of a examinee, it can be the case that CATs are perceived as very difficult for each individual examinee and this could have possible negative side effects, for example, enhanced test anxiety and, consequently, possible lower test performance. This could especially be the case for tests that are administered in primary and secondary education, where, traditionally, tests are constructed in such a way that the average examinee has, on average, a somewhat higher probability (0.60 or 0.70) of correctly answering the items.

For the case of the 1PL and the 2PL model, Eggen (2004, pp.195-219) has proposed a CAT item selection method by which items can be selected with higher or lower difficulty level with only limited loss in measurement efficiency. His method is based on not selecting the item with maximum information at the current θ estimate, but with maximum information at a θ level a bit shifted away from that point to a point at which it has the desired success probability. He compared a CAT with items selection with 0.70 success probability to an unrestricted CAT (0.50 success probability on items), and to a randomly selected test from the bank with 30 items. It was shown that on average the same measurement accuracy was reached with the unrestricted CAT with 11 items and with the easy CAT with 13 items.

Practical Conditions in Item Selection

In the first CATs, the item selection method was based solely on the psychometric criterion of maximum item information. The increasing number of CAT applications has resulted in more consideration being given to content-based and practical requirements or conditions in item selection algorithms. In modern CATs. psychometrically optimal items that meet these practical conditions are selected. Common to these conditions is that they all have a small detrimental effect on the measurement accuracy of the CAT. However, the size of the loss in accuracy generally does not countervail against the practical requirements.

Content Control

When only maximum information selection takes place, this could give results that are in conflict with the desired content specifications of the test. A test constructor could demand that sub-domains of the measured trait are represented in a certain proportion in each CAT. One reason for this demand could be the content and face validity of the test; another could be the requirement to report separate estimates on the sub-domains for diagnostic purposes. An example is that a test on elementary arithmetic should have an equal number of addition and subtraction items. There are several possibilities for realizing such a specification:

1. The item bank is partitioned in sub-banks. For each sub-domain to be distinguished there is a sub-bank and an examinee takes a CAT for each sub-bank. If one uses a variable length for each sub-CAT, there is no complete control on the number of items, and thus on the proportions between the number of items, in the sub-domains. Nevertheless, this approach is often applied. One main reason to do this is that for certain sub-domains sometimes a specific stimulus or item type is used. Alternating items of different sub-domains in one CAT then might cause problems. A CAT on language could, for example, consist of reading and listening items. It might be desirable not to mix these item types during the administration of the CAT.

2. If it is required to have items on sub-domains in fixed proportions in every CAT, it is possible to adapt a CAT algorithm to achieve that. The idea is that the algorithm takes care of the best approximation of the desired specification during the administration of the CAT. An elegant and simple method to achieve this was proposed by Kingsbury and Zara (1991) that operates as follows: After each item, the percentages of items in the sub-domains of all items administered thus far are subtracted from the desired percentages of items in the sub-domains. From the sub-domain that has the highest difference in these percentages, an item that has maximum information is administered next. So in the algorithm, first the sub-domain is determined and within this domain the most informative item is selected.

Exposure Control

In practice, methods that select an optimal CAT have some unwanted properties for certain applications. Although every examinee in principle gets a different test in a CAT, it is commonly found that a few items from the item bank are administered to almost every examinee. Furthermore, a group of items is administered very frequently while other items are hardly ever or never used. Two problems can be identified

1. *Over-exposure:* Some items are used so often that a security risk can occur. If items become known, not only could some examinees profit from this, but in general the item parameters of the items will change as well.

2. *Under-exposure:* Some items are so seldom selected, that one could wonder why they were constructed and kept in the item bank. From a measurement point of view under-exposure might be not a real problem, but for the test constructor or the publisher it could be a major problem.

By putting restrictions on the item selection algorithm, it is possible to try to find a solution for these two problems.

For the over-exposure problem, a number of methods are available and effective at the cost of a little loss in measurement efficiency. In CAT algorithms the item exposure control approach developed by Sympson and Hetter (1985) or generalized forms of it are frequently implemented. In the Sympson-Hetter method, every item has an item exposure parameter specifying the probability that an item is administered once it is selected by the item selection method. In this way it is possible to keep the frequency of administering items below a certain percentage of all CAT administrations. The estimation of the exposure parameters takes place before the testing by simulation studies. Sympson-Hetter is effective against over-exposure of items, but not against under-exposure. Algorithms against under-exposure have been proposed, for example by Revuelta and Ponsada (1998), but they do not always work completely satisfactorily. More equal exposure rates of all the items in the bank almost always go hand in hand with a substantial loss in measurement accuracy. This problem and also all new kinds of exposure control methods are on the research agenda of many CAT researchers.

Research Issues

In this paper, the basics of CAT based on item banks, calibrated with item response theory models, were given. It was shown that the main features of CAT have a sound theoretical basis, but also have been shown to be practical, feasible, and successful. With the help of modern computer technology, valid tests can be constructed and administered that save almost half of the testing time compared to traditional tests. This does not mean that all relevant issues have been solved. Furthermore, educational, technological, and practical developments are a guarantee for the necessity to improve the current possibilities for adaptive testing. Yet several research issues remain.

The development of automatic scoring procedures for items with complex formats will have a major impact on the possibilities for CAT. Richer stimuli will enhance the possibilities for more authentic measurement of competencies and other traits and thus will not only improve the validity of current measurements, but also will offer new applications. As a consequence, the need for IRT models for polytomously scored items in CAT is apparent.

Technological developments will make it possible to maintain item banks on a regular basis. Using the Internet and data warehousing techniques, data from CAT test administrations can be used to keep the item calibration of the bank up to date. Because CAT algorithms always assume that the parameters of the items are known, a major threat to the validity of CATs will be under control. The development of seeding designs and the updating of the estimation and the fitting to the model for the items will need further research.

From the psychometric point of view, major challenges in CAT research will be the development of models and algorithms that take care of multidimensionality of the trait to be measured. In many applications, the trait one wants to measure is described in a multidimensional way. It is, however, not always clear whether the measurement of these dimensions should be conducted simultaneously or separately. In the first case, multidimensional IRT should be applied in CAT. In case more different one-dimensional measurements have to be taken place, searching for the optimal order is an issue. Knowledge of the structure between the different measured constructs can possibly be used for designing optimal CATs.

Another, major issue will be the development of models and CAT algorithms which also take response times to items in account. The easy available measurement of the response times in

CATs make it possible to use them in models for measuring proficiencies and abilities in which there is a speed-accuracy trade-off.

In CAT algorithms, the fulfillment of practical wishes will get more urgent as the number of applications and possibilities for CAT grow. Depending on the major goal of testing, in addition to the algorithms for efficiently estimating θ described in this paper, algorithms for optimal classification of examinees are already available. The same is true for more sophisticated methods for exposure control of the item bank. Especially in application in high-stakes testing environments, the security of the item banks is of utmost importance. Methods for item bank management and exposure control need further development.

References

Clauser, B. E. (Ed). (2002). Advances in computerized scoring of complex item formats. *Applied Measurement in Education, 11* (4).

Dodd, B. G, De Ayala, R. J. & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19*, 5-22.

Eggen, T. J. H. M.. (1990) Innovative procedures in the calibration of measurements scales. In W.H. Schreiber & K. Ingenkamp (eds) . (pp. 199-212). *International developments in large scale assessment.* Windsor, Berkshire: NFER-NELSON.

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. & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees in three categories. *Educational and Psychological Measurement, 60*, 713-734.

Eggen,T. J. H. M. (2004). *Contributions to the theory of practice of computerized adaptive testing.* (Doctoral thesis). Enschede: University of Twente.

Fischer, G. H. & Molenaar, I. W. (Eds.). *Rasch models. Foundations, recent developments, and applications*. New York: Springer-Verlag.

Kingsbury, G. G., & Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4*, 241-261.

Lord, F. M. (1970). Some theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance* (pp.139-183). New York: Harper and Row.

Lord, F. M. (1971. A theoretical study of two-stage testing. *Psychometrika, 36*, 227-242.

Parshall, C. G., Spray, J. A., Kalohn J. C. & Davey, T. (Eds.) (2002). *Practical considerations in computer-based testing.* New York: Springer-Verlag.

Revuelta, J. & Ponsada, V. (1998) A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 38*, 311-327.

Sympson, J. B. & Hetter, R. D. (1985). *Controlling item-exposure rates in computerized adaptive testing.* Paper presented at the annual conference of the Military Testing Association. San Diego.

Van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63,*201-216.

Van der Linden, W. J. & Hambleton, R. K. (Eds.) (1996). *Handbook of modern item response theory.* New-York: Springer-Verlag.

Van der Linden, W .J. & Glas, C. A. W. (Eds.) (2000). *Computerized adaptive testing. Theory and practice.* Dordrecht: Kluwer Academic Publishers.

Van Rijn, P. W., Eggen, T. J. H. M., Hemker, B. T. & Sanders, P. F. (2002). Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 26,* 393-411.

Verhelst, N. D., & Glas, C. A. W. (1995). The one-parameter logistic model. In. G. H. Fischer & I. W. Molenaar (Eds.). *Rasch models: Foundations, recent developments, and applications* (pp.215-237). New York: Springer-Verlag.

Verhelst, N. D., Glas, C. A. W.& Verstralen, H. H. F. M. (1995). One-parameter logistic model (OPLM).[Computer software]. Arnhem: Cito.

Vrabel, M. (2004). Computerized versus paper-and-pencil testing methods for a nursing certification examination: A review of the literature. *CIN: Computers, Informatics, Nursing, 22,* 94-98.

Wainer, H. (Ed.) (2000). *Computerized adaptive testing. A primer.* Second edition. Hilsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H. & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-202.

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