Measurement and Research Department Reports 2001–1

Overexposure and underexposure of items in computerized adaptive testing

T.J.H.M. Eggen

Measurement and Research Department Reports 2001–1

Overexposure and underexposure of items in computerized adaptive testing

T.J.H.M. Eggen

Citogroep
Arnhem, januari 2001

Abstract

Computerized adaptive tests (CATS) have shown to be considerably more efficient than paper-and-pencil tests. This gain is realized by offering each candidate the most informative item from an available item bank on the basis of the results of items that have already been administered. The item selection methods that are used to compose an optimum test for each individual do, however, have a number of drawbacks. Though a CAT generally presents each candidate with a different test, it often occurs that some items from the item bank are administered very frequently while others are never or hardly ever used. These two problems, i.e., overexposure and underexposure of items, can be eliminated by adding further restrictions to the item selection methods. However, this exposure control will affect the efficiency of the CAT. This paper presents a solution for both problems. The functioning of these methods will be illustrated with the results of simulation research that has been carried out to develop adaptive tests.

Introduction

Computerized adaptive tests (CATS) can be used to assess a candidate's general ability in a field or to make placement or pass-fail decisions. It has been shown that these tests are considerably more efficient than paper-and-pencil tests. The literature (see, for instance, Wainer, 2000 and Eggen & Straetmans, 2000) reports that it is possible to either double the accuracy of ability estimation with the same number of items or to halve the average number of items required to take a decision with the same degree of accuracy. This gain is realized by offering each candidate the most informative item from an available item bank on the basis of the results of items that have already been administered.

The item selection methods that are used to compose an optimum test for each individual do, however, have a number of drawbacks. Though a CAT generally presents each candidate with a different test, it often occurs that some items from the item bank are administered very frequently while others are never or hardly ever used. These two problems, i.e., overexposure and underexposure of items, can be eliminated by adding further restrictions to the item selection methods. However, this exposure control will affect the efficiency of the CAT.

After an introduction to adaptive testing, this paper presents a solution for both problems. How these methods work will be illustrated with the results of simulation research that has been carried out to develop these adaptive tests at CITO.

Adaptive tests

Computerized adaptive testing means that individual candidates are tested with the aid of the computer and that the test is constructed while it is being administered: on the basis of results obtained by a candidate, items are administered after they have been selected from an item bank that has been calibrated with the aid of item response theory (IRT).

The item bank

CATs presuppose the availability of an item bank. An item bank is a structured collection of items: besides the item itself, the item bank lists various characteristics of the item. The characteristics of the items that have been filed in the bank may concern content classification, but the most important characteristics in adaptive testing are the items' psychometric features, the item parameters. The algorithms for adaptive tests operate on the basis of these item parameters that are established in calibration research. If this shows that the items meet the requirements of an IRT model and if the item parameters have been estimated with sufficient accuracy, this is called an IRT-calibrated item bank. The IRT model for the item bank that was used for the simulation studies reported on in this paper is the OPLM model (Verhelst & Glas, 1995). In this model, the probability of correctly performing task *i*, also called the item response function, is given by:

$$p_i(\theta) = P(X_i = 1 \mid \theta) = \frac{\exp(a_i(\theta - \beta_i))}{1 + \exp(a_i(\theta - \beta_i))}.$$

Here, β_i is the location parameter of the item. This parameter is associated with the difficulty of the item. This is the point on the ability scale where there is a 50% chance of correctly answering the item. Parameter a_i is the item's discrimination index. Estimates of the values of a_i and β_i for each item have been filed in the item bank.

This paper makes use of a mathematics item bank to illustrate the results and the simulation studies that have been carried out. This item bank consists of 680 items, divided into 4 subdomains: arithmetic (320), information (81), algebra (139), and geometry (140). In the WISCAT (Cito, 1999) test package, this item bank is used for various kinds of adaptive tests developed for vocational education.

The algorithm for administering an adaptive test

An algorithm for an adaptive test contains the regulations for starting, continuing, and terminating the test, and for reporting on a candidate's test performance. Figure 1 is a schematic representation of an adaptive test.

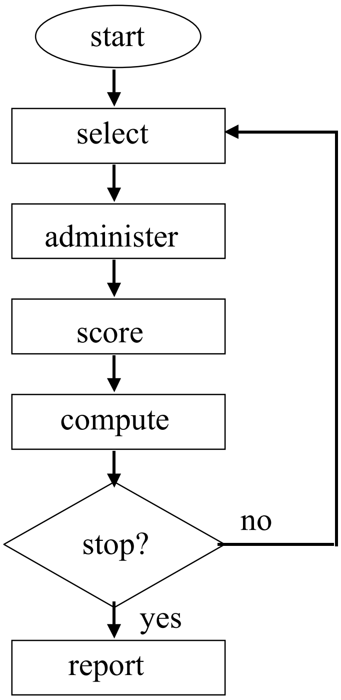

Figure 1. Schematic representation of an adaptive test

As data on the candidate's ability are generally not available before the administration of the test, the test will start with one or more randomly selected items from the item bank. The second step is the item selection that takes place after the administration of each item. An item is selected from the item bank that matches answers given up to that point: the composition of the test is adapted to the candidate. In this way, an optimum test is constructed for each candidate. The criteria for item selection are dealt with in greater detail below; however, what they all have in common is the psychometric notion of information. The underlying idea is that the item that promises to give the most information for the candidate's proven ability up to that point will be administered next. Such optimum item selection is largely responsible for the major efficiency gains of CAT as against a fixed, linear paper-and-pencil test. Efficiency gains means that fewer items are required to assess candidates with the same degree of accuracy, or that they can be assessed much more accurately with the same number of items. Table 1 shows the results of a simulation study for a mathematics placement test. For further results and details, see Eggen & Straetmans (2000).

Table 1. CAT and a linear mathematics intake test

Method	Average number of items	% correct decisions
Paper-and-pencil	25	87,0
CAT (maximum information)	14,2	88,3
CAT (random selection)	20,2	85,2

It is clear that, compared to a paper-and-pencil test, a CAT where items are selected on the basis of maximum information can achieve the same accuracy with less than 60% of the items. If items are randomly selected, there is still an average gain of 20% of the items, but also a loss of accuracy.

After an item has been selected, it is administered and scored. In the subsequent computation phase, the candidate's scores are processed: on the basis of the answer scores on the items, statistical procedures determine the candidate's ability and indicate its accuracy. After each administration of an item, it is decided whether a new item must be selected or whether the administration of the test can be terminated. The criterion for discontinuing the test is generally based on the accuracy with which the candidate's ability has been assessed. If this criterion has not been met, a new item is presented. If the test is discontinued, the administration terminates with a report of the results.

Item selection in adaptive testing

After each item that has been administered, the next item is selected from the item bank. Bearing in mind the objective of the test, an item is selected that best matches the ability demonstrated by the candidate up to that point. Many methods for item selection are available (see, for example, Van der Linden, 1998 and Eggen, 1999). In this paper, the focus is on the following item selection methods, which are also the most widely used in practice: random selection, maximum Fisher information at the current ability estimate and maximum Fisher information at the (nearest) cutting point.

Random selection

If items are randomly selected, this means that, after each administration, each item that has not been used has an equal chance of being selected. However, with such a method of selecting items, the contents of the test are not matched to the ability of the candidate, and testing is not efficient. Simulation studies have shown that random selection eliminates out the possible gains of adaptive testing (see Table 1).

Maximum Fisher information at the current ability estimate

This method of selecting items is optimal if the objective of the adaptive test is to estimate the candidate's ability. In addition, this method also functions fairly well if people need to be assigned to one of two or three categories on the basis of the test. In the OPLM model, the Fisher item information is formally given by:

$$I_i(\theta) = a_i^2 p_i(\theta)(1 - p_i(\theta)) = \frac{a_i^2 \exp(a_i(\theta - \beta_i))}{(1 + \exp(a_i(\theta - \beta_i)))^2}$$

This item information function expresses the contribution an item can make to the accuracy of the measurement of a person as a function of his or her ability. This becomes clear if we realize that the estimation error of the ability estimate can be expressed as a function of the sum of the item information of the items administered: $se(\hat{\theta}_k) = 1 / \sqrt{\sum_{i=1}^{k} I_i(\hat{\theta}_k)}$.

For dichotomous items, the Fisher item information is a single-peaked function of the ability. For example, in the OPLM it shows that, for each item, the information reaches its maximum at the value of the location parameter (difficulty) of the item ($\theta = \beta_i$). In addition, it is clear that the discrimination parameter has a great influence on the information. The information is larger as a_i is larger.

Items are selected with the information function as follows: after the ability estimate $\hat{\theta}_k$ has been determined, the information for each item that has not yet been administered is computed at this point; the item whose information value is highest is then selected and administered.

Maximum Fisher information at the nearest cutting point

With this method, items are ranked on their information at the value of the ability belonging to the (nearest) cutting point for a classification. With classification problems with one cutting point, this method of item selection is better than selection based on maximum Fisher information at the current ability estimate. For problems with two cutting points, no improvement of this method can be expected (Eggen & Straetmans, 2000). One must bear in mind that with this item selection method and one cutting point, each person is presented with the same items in the same order.

Utilization of the item bank

If items are selected according to a psychometric criterion in adaptive tests, then each time the program selects the item that will yield maximum information for the ability demonstrated up to that point. The person's ability and the items' characteristics are optimally attuned. As indicated in Table 1, this yields major gains in measurement efficiency. However, the optimum selection method may have undesirable side effects in practice. This is illustrated in Figure 2, which shows the frequencies of item use from the item bank, also called exposure distribution, for various item selection methods for the placement test (The efficiency gains for each method were reported in Table 1).

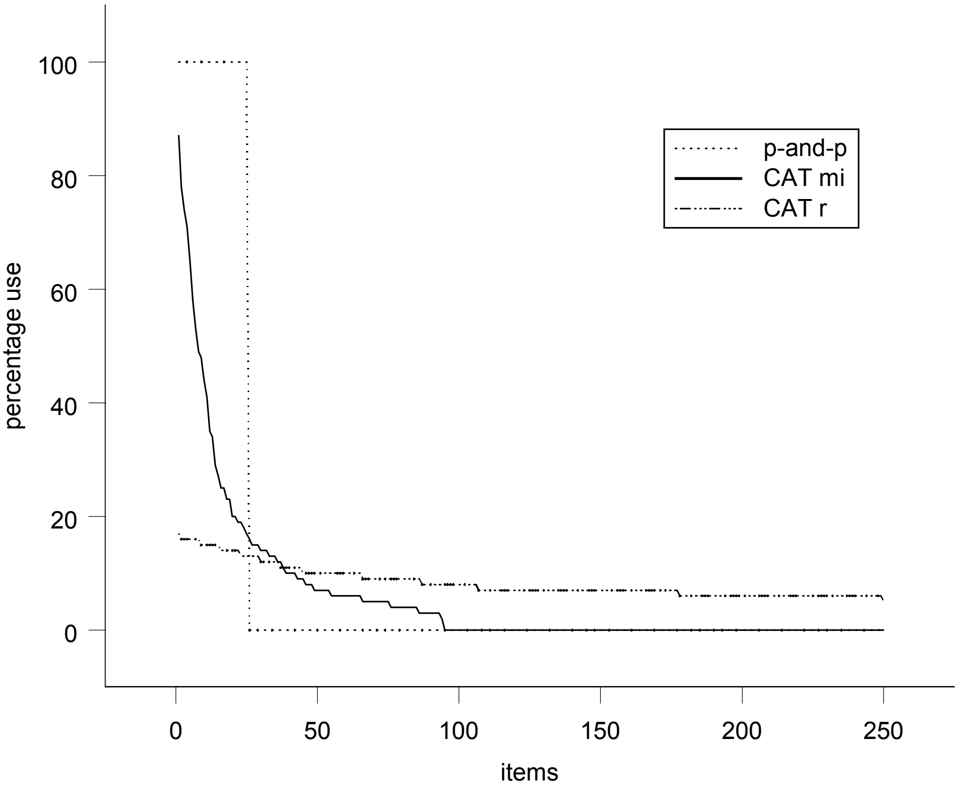

Figure 2. Item utilization a CAT and a linear mathematics placement test

The paper-and-pencil test uses only 25 items from the item bank. The CAT, selecting items on the basis of maximum information, also uses only about a hundred items out of a total of 250 items. In addition, there are some items that are selected very frequently: they are included in the test in well over half the number of test administrations.

Although, in CAT, the optimum item selection method makes sure that virtually each candidate gets a different test, it still happens that some items from the item bank are administered very frequently while others are never or hardly ever used. The major gains in efficiency go together with these two characteristics of the utilization of the item bank, which, in practice, result in the following problems:

1. overexposure: some items are selected so frequently that confidentiality is very rapidly and directly compromised;

2. underexposure: some items are so seldom used that one wonders how the expense of constructing them can be justified.

Figure 2 also shows that these problems do not occur with random item selection: with this selection method, all items are used in 15-20% of test administrations. However, with this method accuracy is lost, and there are few gains compared to the paper version of the test; hence it is unacceptable (see Table 1).

By adding further restrictions to the optimum item selection methods, a solution to these problems can be found, as will be shown below. Imposing preconditions on the algorithm that selects items for maximum information in adaptive testing is the way to meet demands or wishes with regard to the composition of the test. A much-applied restriction, which will not be discussed any further in this paper, is content control. In this case, a tester employs a desirable content specification for the test, establishing that certain components of the ability that is to be assessed occur in a given proportion in each test. For example, an adaptive arithmetic test should contain an equal number of items on percentages and on fractions. For further information on content control, see Kingsbury & Zara (1991) and Eggen & Straetmans (2000).

For problems relating to unbalanced utilization of the item bank, exposure control can be applied. Such exposure control must offer a solution to the underexposure and overexposure of the bank. Imposing restrictions on the optimum test composition will always cause the adaptive test to lose efficiency. Below, exposure control will be discussed in greater detail and a solution presented to both overexposure and underexposure. The extent to which the application of exposure control affects measurement efficiency will also be dealt with.

Exposure control against overexposure

The aim of this kind of exposure control is to reduce the exposure rate of items. This rate is defined as the probability that some item *i* from the item bank will be administered. This probability can be calculated, after a large number of test administrations, by dividing the number of times the item is included in the test by the total number of test administrations. The effect of applying such exposure control should be that items that are administered (to) often without control will be administered less often.

In practice, exposure control proceeds as follows: for each item from the item bank, there is an exposure control parameter (ECP), indicated by k_i, which is a number between 0 and 1. The ECP determines the probability that an item that has been selected with a selection algorithm is actually being administered. There are various methods for determining the k_i's. In the present study the SH method was used, which was originally developed by Sympson & Hetter (1985) and has proved to be very effective. To determine the k_i's by means of the SH method, large numbers of adaptive test administrations (some 1000 per run, for instance) must be simulated according to the desired specifications of the algorithm. These simulation runs must be repeated a number of times.

The SH method

The SH method starts by determining the maximum permissible exposure rate: r, for example 0.2. Then the k_i's are determined in a number of simulation runs. In each simulation run, the selection rate for each item is estimated: $P_i(S)$ and the exposure rate $P_i(A)$. Thus, the number of times an item is selected and the number of times the item is administered is being tracked. In the first simulation run, these rates are equivalent. Step by step, the simulations then proceed as follows:

1. Set all k_i's for item $i = 1,...,I$ to 1.
2. Carry out the simulation and determine $P_i(S)$ and the maximum of $P_i(A)$.
3. Adjust the k_i's for item $i = 1,...,I$:
    a. $P_i(S) > r$, then $k_i = r / P_i(S)$,
    b. $P_i(S) \leq r$, then k_i=1.0,
    c. make sure (by setting the highest to 1) that there are at least as many k_i's equal to 1.0 as the maximum test length.
4. Repeat steps 2 and 3 until the maximum of $P_i(A)$ is approximately equal to r.

Some studies have recently been published in which exposure control methods are compared. The methods focusing on canceling out overexposure are generally variants or extensions of the SH method, such as a SH method with a test length dependency or a SH method that operates conditionally on the ability. On the whole, these methods are more complex, and the results they yielded in comparison with the common SH method did not prompt us to use them.

Exposure control against underexposure

The method described above has proved to be effective against overexposure of the item bank. For underexposure, Revuelta & Ponsoda (1998) have recently proposed an effective method: the so-called progressive method for exposure control. An extension of their proposal is presented here. The basic idea is simple: item selection is based on a mixture of two criteria, viz., chance (R) and maximum information at the current ability estimate (MI). At the start of a test administration, the weight of the R criterion is large and that of the MI criterion is small, but, as the test administration progresses, the weight of R decreases and that of MI increases.

This method is briefly described below. Define:

h: the number of items that have already been administered to a candidate,
m: the maximum number of items a candidate will get,
s=min(a.*h/m*,1): the relative position of an item in the test and
I_i^h: the information of item i at $\hat{\theta}_h$ (the ability estimate after *h* items).

1. Determine the maximum of the information for all unused items after administration of *h* items and call this $H := \max_i I_i^h$.
2. Draw for each item *i* a random number R_i from the uniform division on the interval (0,*H*).
3. Determine for each unused item a weight, a linear combination of a random and an information component: $w_i = (1-s).R_i + s.I_i^h$.
4. Select the item with maximum w_i.

It is easy to see that *s* is a number between 0 and 1 that increases as the length of the test increases. In addition, the definition of w_i shows that the contribution of the random component is large at the start of the test and that item information carries more weight as the test progresses. A special case is chosen, a=1: during the entire test, selection will then be based on a combination of chance and information. If a larger a (for instance 2) is chosen, chance plays less of a role in the item selection (only in the first half of the test, for example). Revuelta & Ponsoda (1998) report that this method with a=1 guarantees that item exposure is spread more evenly over the bank with only a limited loss of accuracy. This presents a solution mainly to the underexposure problem.

The progressive method could be a good supplement to the Sympson-Hetter method of exposure control, which presents a solution to the overexposure problem. The simultaneous application of both methods is expected to solve both the overexposure and the underexposure problems of the item bank.

Simulation studies

In this section, the functioning of the methods proposed for exposure control with the aid of simulation studies that have been carried out with the item bank used in the WISCAT (Cito, 1999) will be illustrated. Only results for one the so-called profile test will be reported. This test decides whether candidates have sufficient mastery of a certain ability level (one cutting point) and concerns all subjects from the item bank (arithmetic, algebra, geometry, and information). As a rule, content control is carried out (an equal number of items for each subject) in the test's algorithm, because the test reports on partial subjects besides deciding on the mastery. As quality indicators of the adaptive tests, the average number of observations required to take a decision (n) and the percentage of correct decisions (%) will be reported. For the utilization of the item bank, the exposure distribution or the percentage of test administrations that include each item will be looked at. In their graphical representations, the items have been consistently ranked from high to low use.

No exposure control

If exposure control is not applied, Figure 3 shows that there are virtually no differences in the quality of the CATs in which items are selected having maximum information at the cutting point (cp) and at current ability estimate (mi). For the exposure division, we do see differences. If items are selected at the cutting point, both the overexposure and the underexposure problems are considerable: a number of items (22) are included in 100% of the adaptive tests, whereas 90% of the total number of 680 items are not used at all. Though in theory this selection method is better than selecting at the current ability estimate, the latter is to be preferred for its item bank utilization alone. In this type of item selection, maximum item use is 68% and there are 19 items that occur in more than 40% of test administrations. The overexposure problem is presented here to.

In addition, it can be seen that only 226 items out of 680 are used. This means that 66% of the items are not used. This was to be expected for some 100 items, which are at too high a level, as the bank is also used to test KSB-3 mastery. Nevertheless, there is major underexposure here as well.

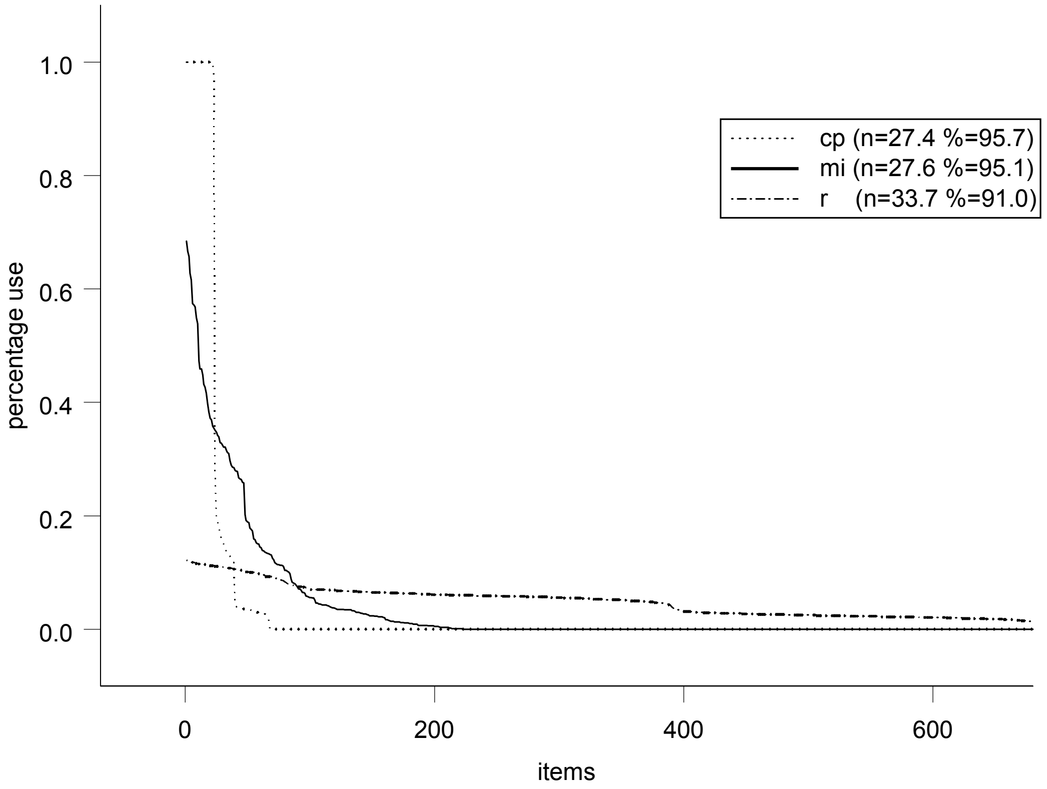

Figure 3. Wiscat item bank utilization without exposure control

Random item selection (r) has also been included in Figure 3. As before, this method does not cause any difficulties with overexposure or underexposure of the item bank. However, the consequences for the quality of the adaptive tests are disastrous. Compared to the other selection methods, both the average number of items needed and the percentage of correct decisions are significantly worse. It should be noted that, for the random sample sizes used in this study, a 0.6 difference in the average number of items is significant at the 99% level, whereas differences in the percentages of correct decisions of 3.3% are significant at this level (2.7% at the 95% level).

The effect of applying exposure control with the SH method

Figure 4 presents the results of applying the SH method, in which the maximum exposure rates have been set at 0.45, 0.35, and 0.25, respectively. With respect to the quality of the tests, it can be seen that, as the maximum exposure rate is set lower, the number of items needed increases somewhat. At a set maximum of 0.25, there is a significant, albeit not very large, increase in the number of items needed compared to selection of items without exposure control. There are virtually no differences in the percentages of correct decisions.

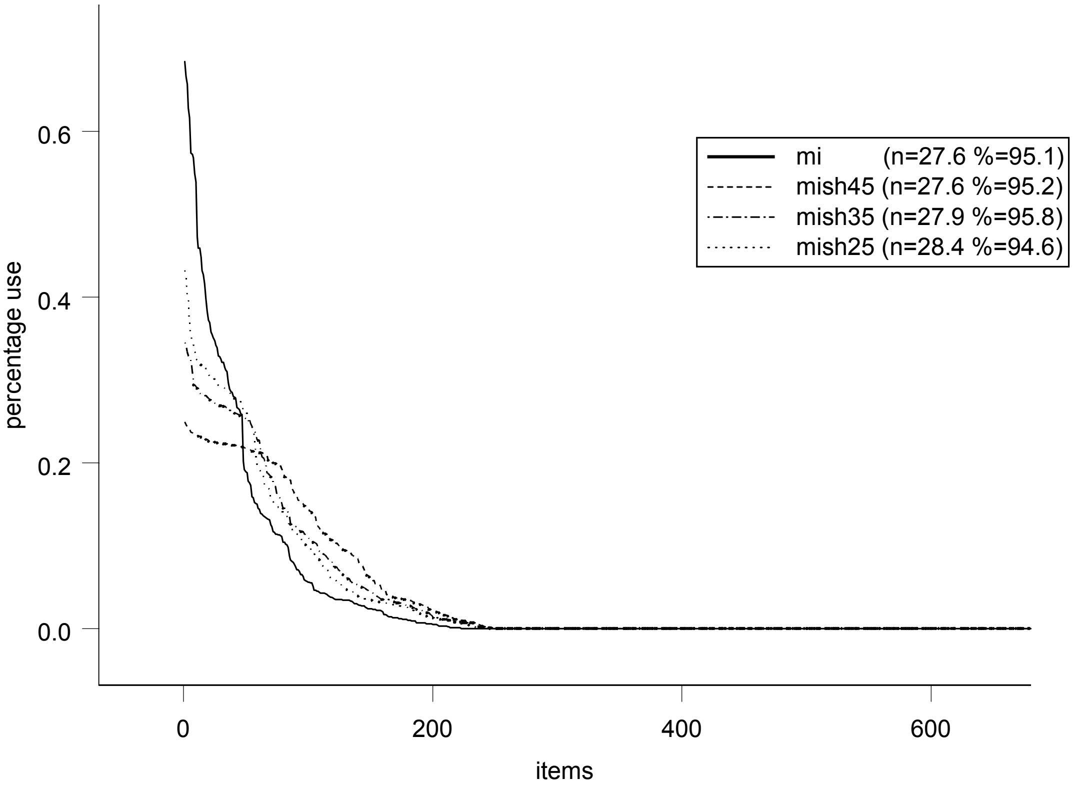

Figure 4: Wiscat item bank utilization applying the SH method for exposure control

As can be seen, the methods are effective against overexposure: the set maxima are not exceeded. However, the SH method does not have much effect on the underexposure of items. For example, with a maximum exposure rate of 0.35, the number of items used has gone up from 222 to 245 compared to selection without exposure control, but 63% of the items from the item bank are still not being called upon.

The effect of applying exposure control with the progressive method

Figure 5 shows the results of applying the progressive method with a=1, 2, and 3, respectively. Item selection is based on a combination of information and chance during the entire adaptive test, only the first half, and only in the third part of the test, respectively. It can be seen that applying this kind of exposure control leads to hardly any quality differences in the adaptive tests. Only if the progressive method is applied during the entire test a slight increase in the average number of items needed is observed.

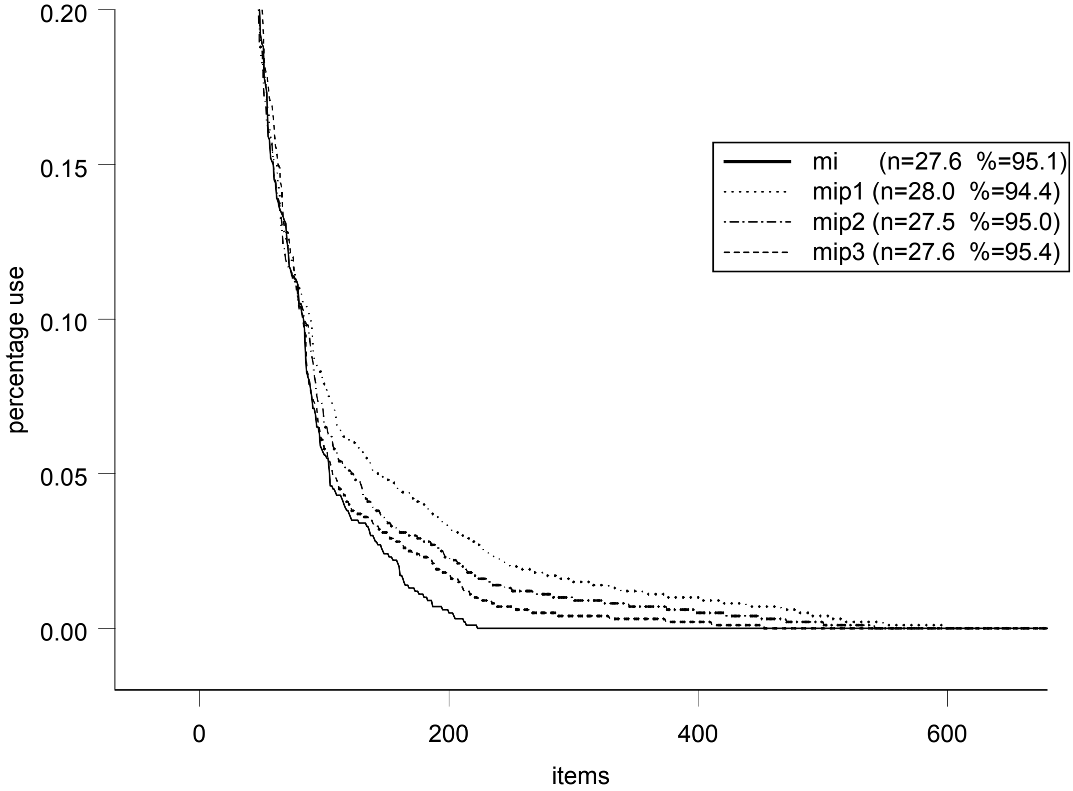

Figure 5: Wiscat item bank utilization applying the progressive method for exposure control

It is clearly observable that the progressive method is effective against the problem of underexposure: at a=1, 591 items are used; at a=2, a total of 530 items are used; and at a=3, 510 items. So only between 15-25% of items remain unused. The progressive method is not effective against overexposure: at a=1, 2, and 3, the maximum use of an item is 55.8%, 60.2%, and 62.8% of administrations, respectively.

The effect of applying the SH method together with the progressive method

Figure 6 presents the results of applying the SH method with a maximum exposure rate of 0.35 in combination with the progressive method with a=1, 2, and 3. The 0.35 boundary was chosen because it was acceptable for intrinsic considerations and because no significant deterioration of quality has been observed compared to selection without exposure control (see Figure 4). When these two methods are combined, first the k_i parameters of the SH method are established in advance, and subsequently, these parameters and the progressive method are applied during testing. A look at the results shows that applying this combination of methods leads to an average increase in the number of items needed of approximately 1. The percentages of correct decisions remain virtually the same.

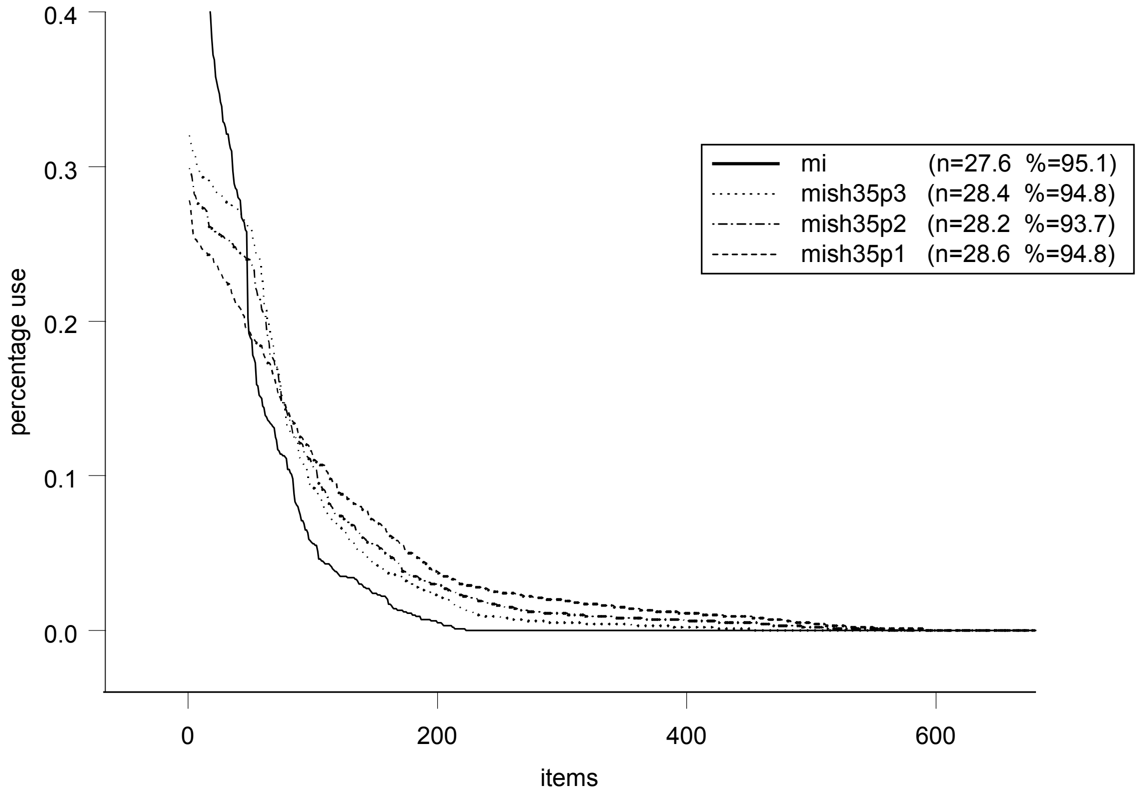

Figure 6: Wiscat item bank utilization combining the progressive method and the SH method

The effect on item bank utilization is favorable: the combination can cope with both the underexposure and the overexposure problems. With the SH method in combination with the progressive method with a=1, 2, and 3, the numbers of items used are 604, 543, and 525, respectively. The maximum exposure rates are 0.28, 0.31, and 0.32, which is consistently below the set maximum of 0.35. Thus, the methods, jointly applied, are effective against the problems, but at the expense of an average of 1 item in test length. This difference is statistically significant compared to selection without exposure control, but, for practical purposes, this is an acceptable loss of measurement efficiency.

On the basis of these results, exposure control in the definitive algorithm of this Wiscat adaptive test will take place combining the SH method with a maximum exposure rate of 0.35 and the progressive method with a=2.

Conclusion

Computerized adaptive tests (CAT) can achieve considerable efficiency gains compared to traditional tests. These gains are realized by using item selection methods that put together an optimum test for each person. This paper showed that the optimum item selection methods also have some less desirable features in practice. Although, with CAT, each candidate gets a different test, it is frequently the case that some items from the item bank are selected very often while another part of the item bank is never or seldom used at all. Too frequent use may compromise the confidentiality of the items, which is a major practical problem in testing. If items are never or hardly ever used at all, this might be considered a waste of money and energy spent on item construction.

This paper presented solutions for both the problem of overexposure and the problem of underexposure. The functioning of the methods and the consequences for testing efficiency were illustrated with the results of research carried out to develop these tests. This paper presented the results of only one test, but replication of the simulations for other tests with the same item bank and also with other item banks consistently yielded comparable results.

It is concluded that the SH method presents an effective solution to the overexposure problem. The progressive method is effective against underexposure. Combined application addresses both problems. The restrictions imposed on item selection by exposure control can easily be included in the algorithms for adaptive tests.

References

Cito. (1999). *WISCAT. Een computergestuurd toetspakket voor rekenen en wiskunde.* [A computerized test package for arithmetic and mathematics]. Cito: Arnhem.

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 into three categories. *Educational and Psychological measurement, 60*, 713-734.

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.

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 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.

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

Wainer, H., Dorans, N.J., Flaughter, R., Green, B.F., Mislevy, R.J., Steinberg, L. & Thissen, D. (2000). *Computerized adaptive testing: A Primer. Second Edition.* Hillsdale (NJ): Lawrence Erlbaum.