Optimal testing with easy items in computerised adaptive testing

Theo Eggen and Angela Verschoor

Citogroep Arnhem

Paper IAEA Conference, Manchester, UK ,October 2003

Introduction

Computerized adaptive tests (CATs) are individualized tests that are administered in an automated environment. CATs are used for estimating the ability of a student or for making a decision on, for instance, the most appropriate training program for that student. It has been shown that, compared to traditional linear tests CATs yield a considerable gain in efficiency. In the literature (see, e.g., Wainer, 1990 and Eggen & Straetmans, 2000), it has been reported that halving the average number of items needed is feasible, while at the same time the accuracy of the ability estimate or the decisions taken is maintained. CATs make use of item banks which are calibrated using item response theory (IRT) (Hambleton & Swaminathan, 1985). The gain in CATs is realised by selecting, on the basis of the results on previously administered items, the most informative item from an available item bank. During testing, the optimal item is chosen after every item for every student and thus the optimal test is assembled and administered.

CAT-tailored testing has a number of frequently mentioned advantages: the gain in measurement efficiency goes hand in hand with the fact that each student is challenged at his or her own level because items which are too difficult or too easy for a given student will never be administered. Initially, the intended optimality, and by that the item selection method, was based solely on a measurement theoretic or psychometric criterion. The criterion of maximum item information at the current ability estimate is in common use (Van der Linden & Pashley, 2000). The increasing number of CAT applications has resulted in more consideration giving to content-based and practical requirements or conditions in item selection algorithms. Applying content control (Kingsbury & Zara, 1991) and exposure control (Eggen, 2001) is routinely possible. In modern CATs, items that are psychometrically optimal are selected from an item bank in which is care taken to fulfil these practical conditions.

The aim of the present study was to determine whether it is possible to come towards tested persons even more by not only fulfilling the practical conditions, but also by releasing the psychometrically optimal selection. This is because that psychometrically optimal selection of items has the consequence, that for an individual student, items will always be chosen which he or she, at his or her thus far known ability level, has a 50% probability of answering correctly. Thus, as a rule, students taking a CAT will answer about only half of the items correctly. Although in the scoring of a student the difficulty of the items is taken into account, it can be the case that CAT tests are perceived as very difficult for each individual student and this could have possible negative side effects, for example, enhanced test anxiety and, consequently a possible lower test performance. This could especially be the case for tests which are administered in primary and secondary education, where traditionally tests are constructed in such a way that the average student has, on average, a somewhat higher probability (60 or 70%) of correctly answering the items.

One approach for reducing possible negative effects on the difficulty of the items is self-adaptive testing (Rocklin & O' Donnell, 1987}. Self-adapted tests (SATs) are CATs in which the difficulty level of each item is chosen by the examinee rather than by the CAT algorithm. SATs have been studied rather extensively in recent years. The meta-analysis by Pitkin & Vispoel (2001), comparing SATs with CATs, gives an overview. In general, test anxiety reduction is reported in SATs and there is also a little gain in the average performance of examinees if SATs are compared to CATs. The reduction of anxiety as well as a possible larger bias in the ability estimates in the SATs as a consequence of self-selection of the items are possible explanations. Compared to CATs, SATs are less reliable: more items are needed to reach the same measurement precision. Finally, it can be mentioned that, for students a SAT is more time consuming to take a sat than a CAT, and that implementing a sat in practice still has a number of issues that are not completely clear.

In the present study the possibilities for using CATs with selection methods in their algorithms which lead to higher (or lower) success probabilities than 50% were explored. Changing the CAT algorithm for that reason was also proposed in a study by Bergstrom, Lunz & Gershon (1992). They successfully applied an algorithm which chooses easier items, but only for the case of the one-parameter logistic IRT model. In this study, for both the one- and the two-parameter logistic IRT model, two CAT selection methods, choosing items with varying difficulties, were developed and the consequences for the measurement efficiency evaluated.

Item selection in CAT

Computer adaptive tests presuppose the availability of an IRT-calibrated item bank. The algorithms for adaptive tests operate on the basis of the item parameters from an IRT model. The IRT model used in this study is the two-parameter logistic model (2pl). In this model, the probability of correctly answering item i, also called the item response function, is given by

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

Here, β_i is the location parameter of the item. This parameter is associated with the difficulty of the item. It is the point on the ability scale at which the student has a 50% chance of correctly answering the item. Parameter α_i is the item's discrimination parameter. In case the discrimination parameter for all items is the same $\alpha_i = a$, we have the special case of the one-parameter logistic model (1pl). In a calibrated item bank, estimates of the values (of α_i and) β_i for each item have been filed in the bank. After each item has been administered, the next item is selected from the item bank. An item is selected that best matches the ability demonstrated by the candidate up to that point.

Usually the (Fisher item) information is used for selecting. In the case of the two-parameter model, this function is given by

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

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

Items are selected according to the following procedure: 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.

The item information function

For dichotomous items, the Fisher item information is a single-peaked function of the ability. In the two-parameter model, 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.

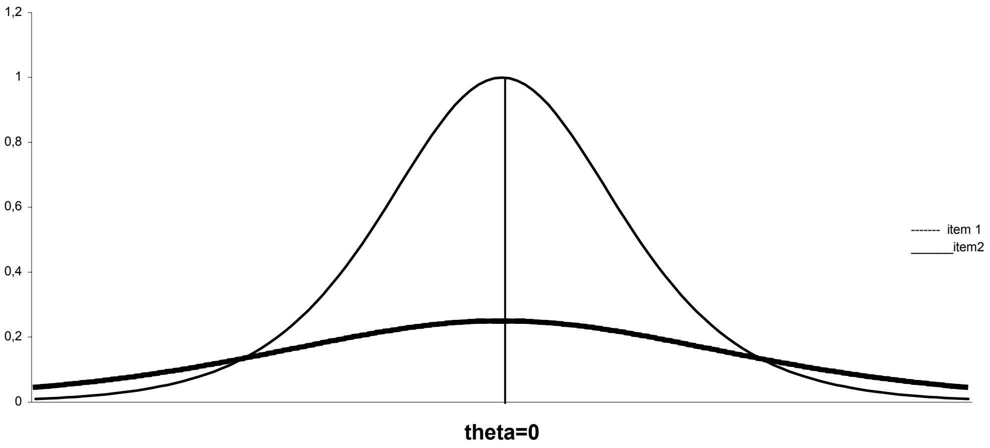
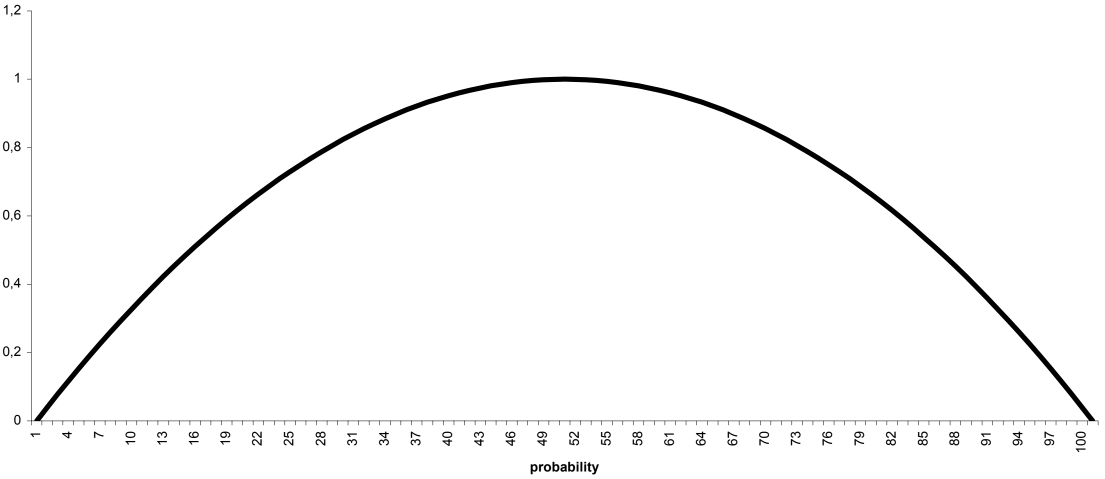


Figure 1. Item information functions : $\beta_1 = \beta_2 = 0$ and $\alpha_1 = 1, \alpha_2 = 2$

Figure2. Relation success probabilty and information



The relation between the information in an item and the probability of succeeding on an item in the 1pl and the 2pl is given in Figure 2.

One can see that an item gives maximum information at success probability of 0.50. At other probability levels, there is always less information.

Item selection on the basis of success probability

For each item, ability levels can be defined at which there is a certain success probability on an item. This is what we call the probability points of an item. For instance, the p-60 point of an item is the ability level at which there is a probability of 0.60 to answer the item correctly. The p-points are easily determined.

Consider the probability of correctly answering an item

$$P_i(\theta) = \frac{\exp \alpha_i(\theta - \beta_i)}{1 + \exp \alpha_i(\theta - \beta_i)} \ .$$

For a given probability, the ability pertaining to that point is then determined from:

$$\ln \frac{P_i(\theta)}{1 - P_i(\theta)} = \alpha_i(\theta - \beta_i) \ ,$$

from which it follows that

$$\theta = \beta_i + \frac{1}{\alpha_i} \ln \frac{P_i(\theta)}{1 - P_i(\theta)} \ .$$

Then the p-x point (with a probability of x) of an item is defined as

$$p_i\_x = \beta_i + \frac{1}{\alpha_i} \ln \frac{x}{1 - x} .$$

It is easily seen that the p-50 point of an item equals the difficulty parameter β_i.

If the item selection in a CAT takes places on the basis of success probability, this can be achieved as follows. Select the item for which the distance between the current ability estimate and the p_i_x point is minimal:

$$\min_{i \in items} \left| \hat{\theta} - p_i\_x \right| .$$

Performance of item selection based on nearest p-point

To evaluate the performance of the item selection methods, simulation studies were conducted. First, the results of a simulation study with an item bank calibrated with the 1pl will be given, followed by a study with an item bank calibrated with the 2pl.

The one-parameter model item bank

The 1pl item bank consists of 300 items with $\beta \sim$ N(0,1). The CAT algorithm used starts with an item of intermediate difficulty (one item randomly selected from 114 items with $-0.5 \leq \beta \leq 0.5$) and has a fixed test length of 40 items. In the simulation, samples of 4000 abilities were drawn from the normal distribution : θ ~N (0,1). The selection methods at the different success probabilities were compared. As baselines in the comparison, the simulations are also conducted with random selection of all items, and the optimal maximum information selection at the current ability estimate. The results of the simulations are given in Table 1.

Table 1: simulation 1pl CAT: selection nearest p-point.

Selection method	Mean error $1/n\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)$	mean $se(\hat{\theta}_k)$ (sd)	mean % correct (sd)
Max info	0.006	0.328 (0.015)	49.7 (8.6)
P_10	0.041	0.435 (0.009)	22.4 (14.5)
P_20	0.048	0.384 (0.045)	27.3 (12.4)
P_30	0.035	0.352 (0.024)	33.5 (10.3)
P_40	0.016	0.334 (0.015)	41.1 (8.8)
P_50	-0.013	0.328 (0.012)	50.0 (8.5)
P_60	-0.016	0.333 (0.017)	58.1 (9.2)
P_70	-0.029	0.351 (0.024)	65.4 (11.0)
P_80	-0.043	0.379 (0.044)	71.4 (13.5)
P_90	-0.034	0.424 (0.098)	75.2 (15.6)
Random	0.007	0.383 (0.078)	50.0 (19.9)

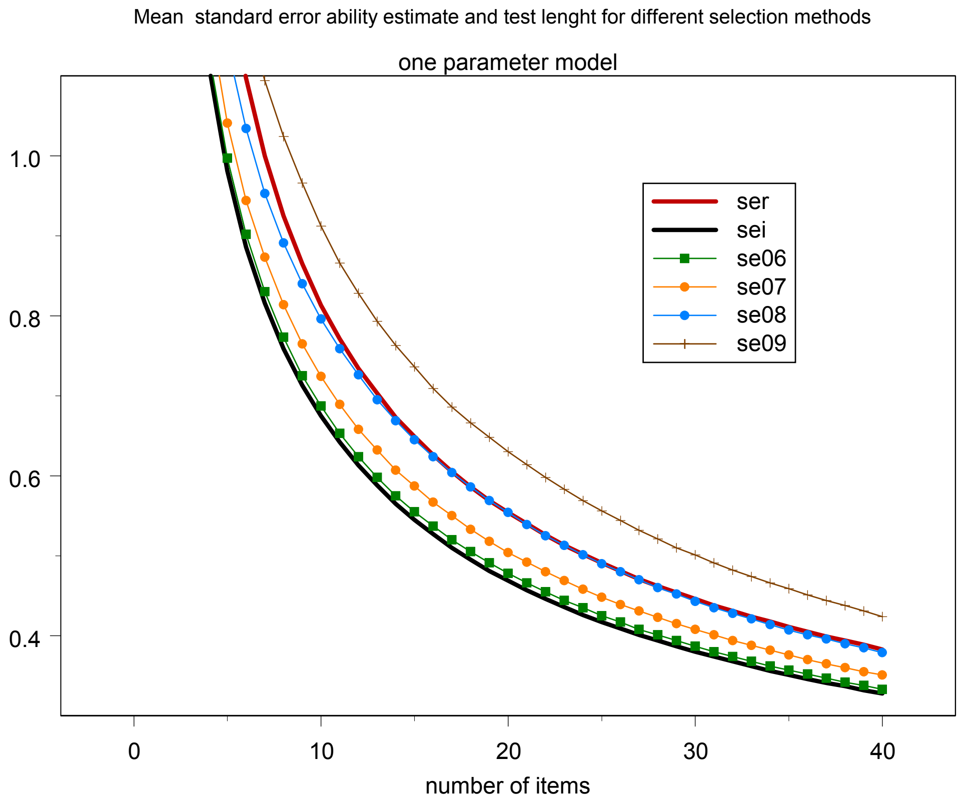
First it is noted that there is, on average, a small discrepancy between the known abilities and the estimated abilities, and the effect seems to be systematic: when the items with a success probability lower than 0.50 are chosen, there is a overestimation of the ability; when selection takes places with higher success probabilities, the ability is generally slightly underestimated. This effect is in line with the known small bias of the ability estimator used (Warm, 1989), but the effect is opposite to the bias in the maximum likelihood estimator of the ability as was reported in Pitkin & Vispoel (2001) in case a test is not optimally assembled at an ability level.

The selection methods show in the results on the percentages correct an effect in the desired direction. Selecting at a success probability higher or lower than 0.50 does not necessarily lead to the same percentage of correct answers of the simulated examinees. The more extreme the probability is, the larger the discrepancy between the selection percentage and the percentage correct. This can be explained by the fact that only a limited number of extremely difficult and extremely easy items are available in the item bank.

If we look at the mean standard errors of the ability estimates with the selection methods, the expected effect can be seen. In the 1pl model, selection at maximum information is equivalent with selection of the item at the nearest p-50 point. Non-optimal selection, at other success probabilities, has an expected negative effect on measurement precision. The effect with the current item bank is symmetric around the p-50 point selection: selection at the nearest p-(50+x) point leads to about the same loss in precision as selection at the nearest p-(50-x) point.

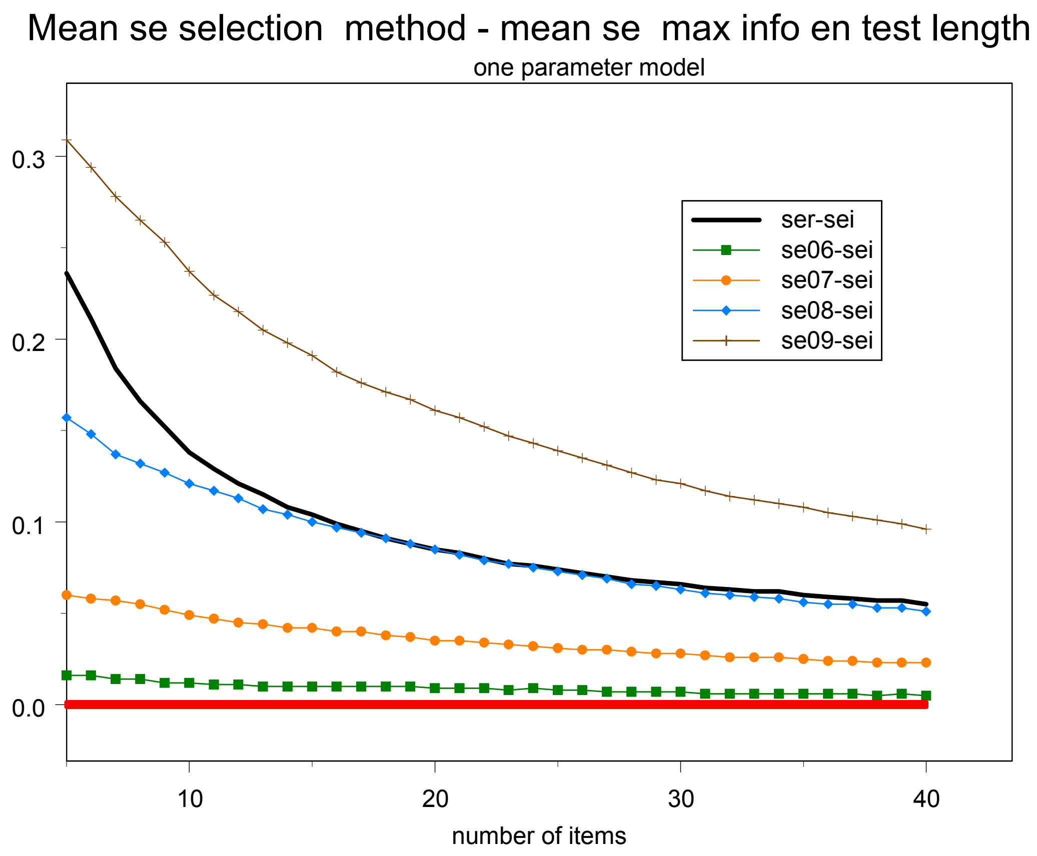
The performance of selection methods can be compared better if the mean standard errors are considered as a function of the test length. In the next figures the results for the selection methods with a success probability of higher than 50% are plotted.

Figure 3


Mean standard error ability estimate and test lenght for different selection methods
one parameter model
1.0
0.8
0.6
0.4
0
10
20
30
40
number of items
ser
sei
se06
se07
se08
se09

3.

Figure 4



It can be seen that the easier the selected items, the greater the loss in measurement precision is. The loss in precision when items are selected at the nearest p-60 and p-70 point is rather small. Selecting at p-80 is as bad as random selection, while selection at p-90 is even worse.

Table 2 gives the number of items needed on average with a selection method to achieve measurement precision which is equivalent with a test of 30 randomly drawn items from the bank.

Table 2. 1pl bank; selection on p-points; equivalence with random test of 30 items

Selection method	Number of items
Max info	22
p-60	23
p-70	25
p-80	30
p-90	37

The two-parameter model item bank

The 2pl item bank consists of 300 items with $\beta \sim$ N(0,0.35) and $\ln\alpha : N(1,0.3)$ The CAT algorithm used starts with an item of intermediate difficulty (one item randomly selected from 113 items with $-0.17 \leq \beta \leq 0.17$) and has a fixed test length of 40 items. In the simulation, samples of 4000 abilities were drawn from the normal distribution : θ ~N (0,0.35). The selection methods at the different success probabilities are compared in Table 3.

Table 3: simulation 2pl CAT: selection nearest p-point.

Selection method	Mean error $1/n\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)$	mean $se(\hat{\theta}_k)$ (sd)	mean % correct (sd)
Max info	0.001	0.085 (0.013)	49.0 (11.9)
P_10	0.011	0.117 (0.033)	26.5 (18.5)
P_20	0.008	0.116 (0.020)	28.5 (14.1)
P_30	0.005	0.114 (0.013)	34.1 (10.6)
P_40	0.001	0.110 (0.010)	41.6 (8.9)
P_50	0.001	0.111 (0.008)	49.7 (8.5)
P_60	-0.009	0.110 (0.009)	58.0 (9.4)
P_70	-0.006	0.115 (0.015)	64.9 (11.8)
P_80	-0.009	0.114 (0.018)	70.8 (15.0)
P_90	-0.012	0.124 (0.033)	74.5 (17.6)
Random	0.001	0.132 (0.033)	49.7 (19.5)

The results for the mean percentages correct are about the same as in the case of the 1pl item bank. The same is true for the systematic bias in the ability estimates, although there is hardly any bias. The results on measurement precision show that selecting on higher or lower p-points has a very negative impact compared to maximum information selection. One sees that the more extreme the success probabilities are, the larger the loss in precision is , but in any case the loss is considerable, which is even clearer from Figure 5.

Figure 5

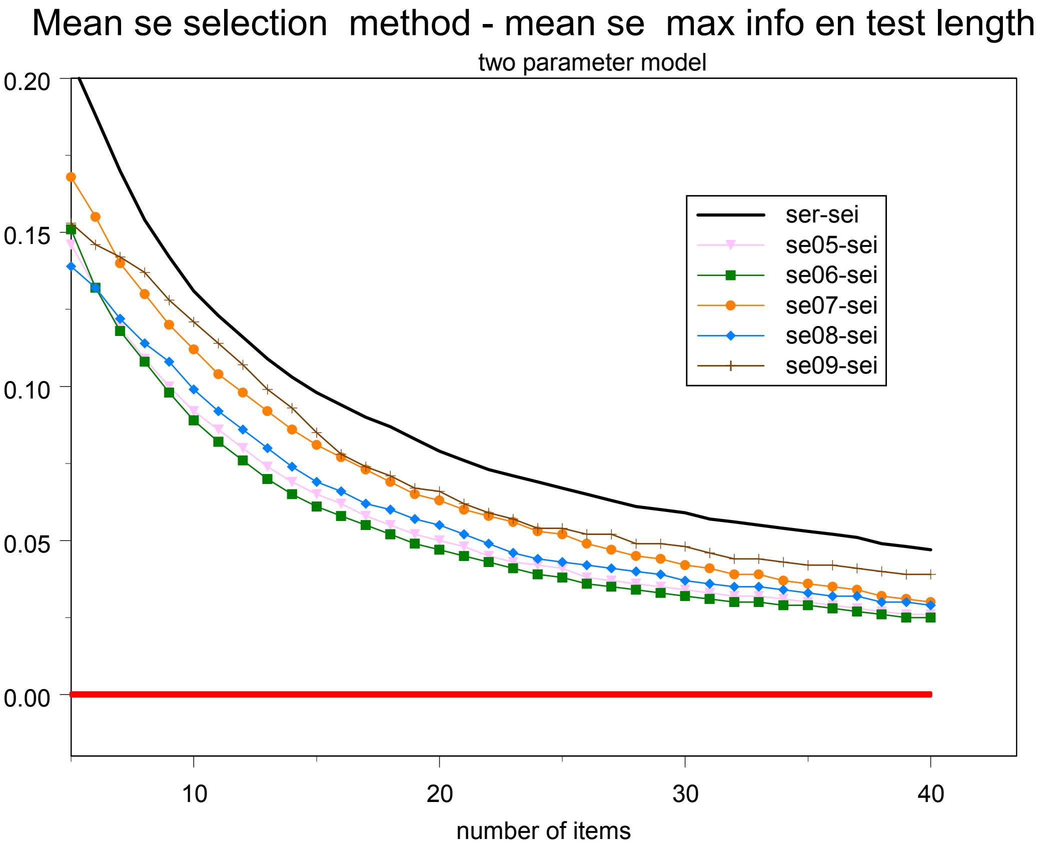


Table 4 gives the number of items needed on average with a selection method to get measurement precision which is equivalent with test of 30 randomly drawn items from the bank.

Table 4. 2pl bank; selection on p-points; equivalence with random test of 30 items

Selection method	Number of items
Max info	10
p-50	22
p-60	21
p-70	25
p-80	23
p-90	26

It is seen that selecting at the nearest p-60 point doubles the number of items needed compared to maximum information selection. So, selecting at the nearest p-point of an item works quite well in an item bank based on the 1pl model, but for item banks calibrated with the 2pl model, the results are very poor.

Alternative method for selecting with higher or lower success probabilities.

The problem encountered with selection on success probability is due to the fact that, in selection, only the success probability of an item is considered, but not the values of the information function of the items. In the 1pl model, this has no consequences owing to the fact that, in that case, all information functions have the same shape; they only differ in the point where they reach their maximum ($\theta = \beta_i$). This implies that the differences between, for instance, a p-50 point and a p-60 point of the item is constant for every item. In the 2pl model, however, the (absolute) value of the information plays an important role. Neglecting this and selecting only on success probability has negative consequences.

What goes wrong is illustrated in Figure 6 which shows the item response curves and the information functions of two items with the same difficulty but with different discrimination parameters. For the first item, the discrimination parameter is $\alpha_i = 1$ (dotted curves); for the second item, $\alpha_i = 2$. On the ability axis at point 1, the coinciding p-50 point of both items is given; at point 2 and 3, the p-60 point of item 2 and item 1 respectively. If we select items at the nearest p-50 point, we can see that if the current ability estimate is at point 1, for this method, both items could be chosen, while at this point the information in item 2 is much higher. Another example: if we select at the nearest p-60 point and the current ability estimate is at the indicated arrow or higher, item 1 is preferred, while the information is much higher for item 2.

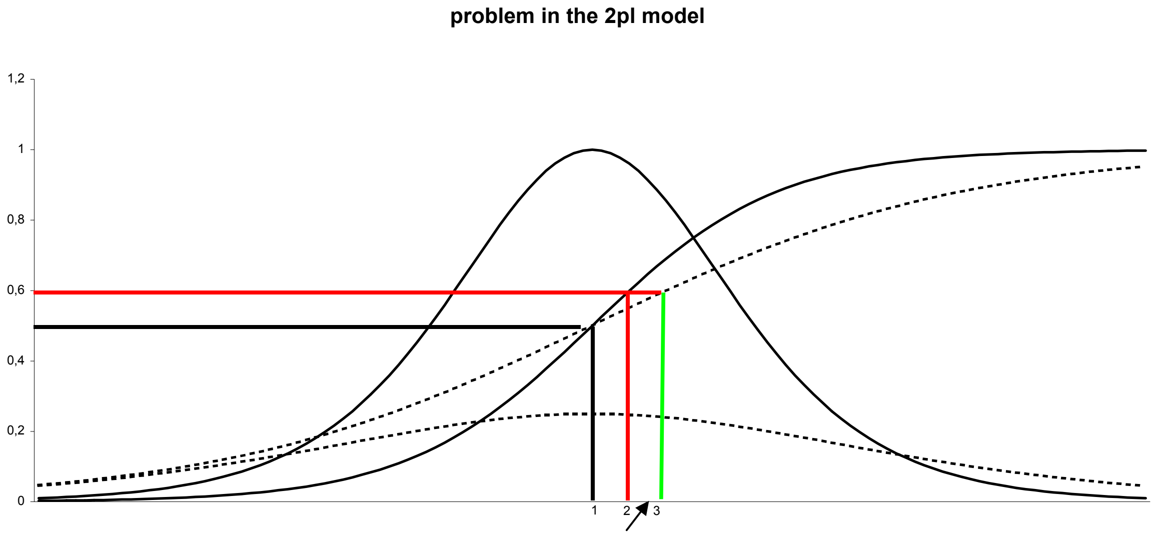


Figure 6: Item response curves and information function of two item in 2pl.

In order to overcome this problem, a new selection method was developed which takes account of the success probability and of the absolute value of the information. The idea is not selecting items with maximum information at the current ability estimate, but selecting the item with maximum information at a lower or a higher level of the ability than the current ability estimate. If easier items (with higher success probabilities) are wanted, one chooses items which are optimal (have max info) at an ability level which is below the current ability estimate. If harder items are desired, the items are selected at an ability point above the current estimate.

Suppose the current ability estimate is $\hat{\theta}$, then easier or harder items are selected by searching at an ability level of $\hat{\theta} - y$,with y positive for easier items and negative for harder items. The value of the shift on the ability can be deduced from the desired success probability. In the 2pl model, it yields

$$P_i(\theta) = \frac{\exp \alpha_i(\theta + y - \beta_i)}{1 + \exp \alpha_i(\theta + y - \beta_i)} \quad .$$

From that it follows

$$\alpha_i(\theta + y - \beta_i) = \ln \frac{P_i(\theta)}{1 - P_i(\theta)} \quad .$$

And in order to get a certain success probability, the shift on the scale is

$$y = \frac{1}{\alpha_i} \ln \frac{P_i}{1 - P_i} \quad .$$

So, for selecting items with a desired success probability of 60%, items are selected which have maximum information at

$$\theta = \hat{\theta} - \frac{1}{\alpha_i} \ln 1.5 \quad .$$

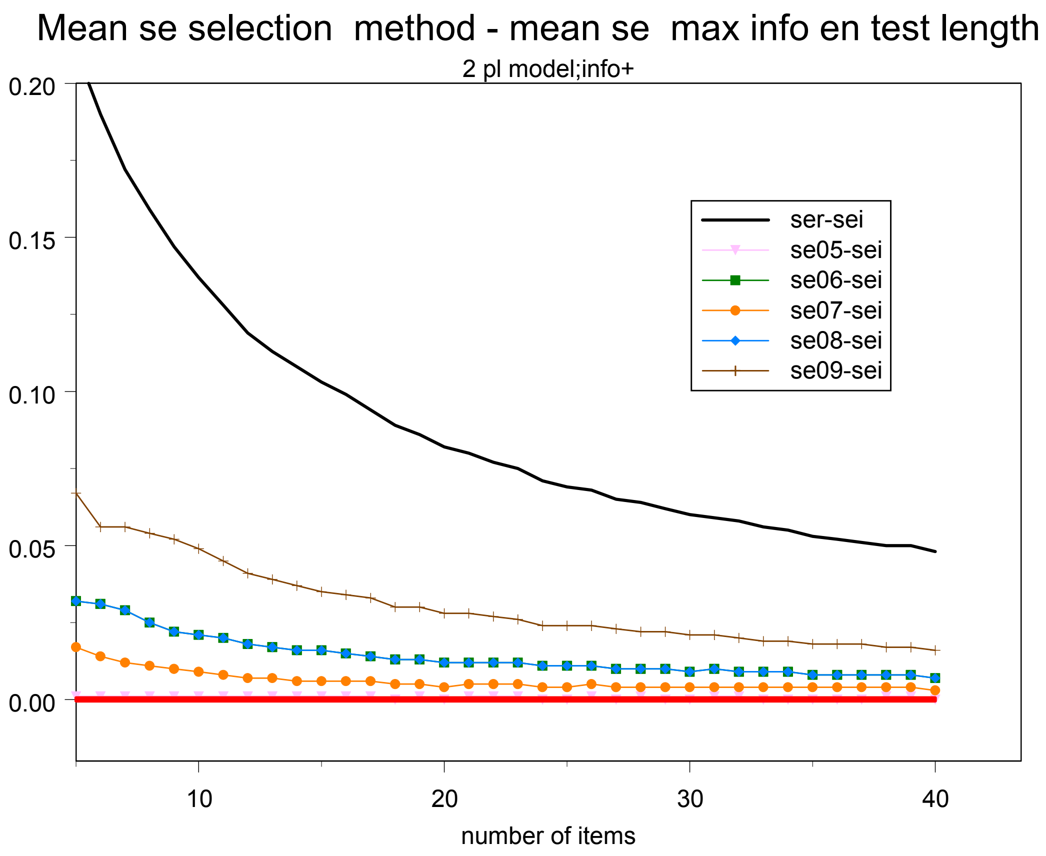
In the one-parameter model, this selection method is equivalent to selecting items on p-points nearest to the current ability estimate. In the two-parameter model, however the selection is quite different. Applying the selection methods in the simulation studies on the same 2pl item bank gives the results in Table 5.

Table 5: simulation 2pl CAT: selection at shifted ability level

Selection method	Mean error $1/n\sum_{i=1}^{n}(\hat{\theta}_i-\theta_i)$	mean $se(\hat{\theta}_k)$ (sd)	mean % correct (sd)
Max info	0.001	0.085 (0.011)	49.0 (12.2)
P_10	0.010	0.100 (0.018)	28.6 (16.1)
P-20	0.006	0.091 (0.012)	33.5(14.2)
P_30	0.004	0.088 (0.012)	38.8(13.4)
P_40	0.001	0.086 (0.013)	43.8(12.6)
P_50	-0.003	0.085 (0.013)	49.4(12.3)
P_60	-0.004	0.085 (0.012)	55.1(12.2)
P_70	-0.002	0.088 (0.012)	60.9(12.6)
P_80	-0.008	0.092 (0.015)	65.4(14.1)
P_90	-0.011	0.101 (0.015)	71.7(15.6)
Random	0.003	0.133 (0.031)	50.5(19.6)

The results for the mean % correct are about the same as with selecting on nearest distance to p-points. The same is true for the systematic bias in the ability estimates, although there is hardly any bias with the new selection method. The results on measurement precision show that selecting easier or harder items is possible with the new selection method without a large loss in precision. This result is also seen from Figure 7.

Figure 7



It is clear that selecting easier or harder items with the new selection method does not cause much loss in measurement precision. If one aims at a success probability of 60%, there is hardly any loss; the more the items are chosen, the larger the loss in efficiency. But at all success probabilities, the random selection is far outperformed in contrast to the results with the selection on the p-points. This result will become clearer in Table 6.

It gives the number of items needed on average with a selection method to get a measurement precision which is equivalent with a test of 30 randomly drawn items from the bank.

Table 6. 2pl bank; selection shifted ability level; equivalence with random test of 30 items

Selection method	Number of items
Max info	10
p-50	10
p-60	10
p-70	11
p-80	12
p-90	16

The new selection method seems to perform without any large loss in measurement precision: with the current item bank and algorithm it is possible to reach a percentage correct of 60% at the cost of, on average, 1 item compared to the optimal test.

Simulation with item selection applying exposure control

Because standard some form of exposure control is applied in the selection algorithm in modern CATs, it was investigated whether the new algorithm still works when exposure control is added to the CAT algorithm. The results of the same simulations, but combined with the application of the Sympson-Hetter exposure control with an maximum exposure of 0.3 (see e.g. Eggen, 2001), are given in Tables 7 and 8.

Table 7: simulation 2pl CAT: selection at shifted ability level and exposure control

Selection with SH 0.3	Mean error $1/n\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)$	Mean $se(\hat{\theta}_k)$ (sd)	mean % correct (sd)
Max info	0.002	0.098 (0.010)	49.0 (10.0)
P_50	0.001	0.098 (0.008)	49.5(10.0)
P_60	0.002	0.100 (0.011)	54.5(11.4)
P_70	-0.017	0.104 (0.013)	55.5(12.8)
P_80	-0.008	0.106 (0.011)	59.3(14.4)
P_90	-0.005	0.111 (0.016)	60.0(15.6)
Random	0.003	0.133 (0.031)	50.5(19.6)

Again there is hardly any bias and the differences in the percentages correct seem to be less than in selecting without exposure control. The discrepancy between the desired and the achieved percentages correct is larger when exposure control is applied. With respect to measurement precision, the results are similar to selecting without exposure control. The number of items needed to get an equivalent to a test with 30 randomly selected items is given in Table 8. It is clear that applying exposure control on average costs 2 or 3 items .

Table 8. 2pl bank; selection shifted ability level and exposure control; equivalence with random test of 30 items

Selection method with SH =0.3	Number of items
Max info	12
p-50	12
p-60	13
p-70	14
p-80	15
p-90	18

Simulations with an infinitely large item bank.

A possible explanation for this discrepancy between the desired and the achieved percentages correct is that there is a mismatch between the items available in the item bank and the desired percentages in the population. One possible solution for this could be enlarging the size of the item bank. In order to check this, simulations were conducted with a very large item bank.

The 2pl item bank consists of 3000 items, 1000 with α =2, α =3 and α =4 and the difficulty parameter from a uniform distribution $\beta \sim$ U(-1.1,1.1) The CAT algorithm used starts with an item of intermediate difficulty and has a fixed test length of 40 items. In the simulation, samples of 4000 abilities were drawn from the normal distribution : $\theta \sim$N (0,0.35). The selection methods for different success probabilities are compared in Tables 9 and 10.

Table 9: simulation 2pl CAT large item bank; selection at shifted ability level.

Selection	Mean error $1/n\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)$	Mean $se(\hat{\theta}_k)$ (sd)	mean % correct sd)
Max info	-0.001	0.083 (0.002)	50.1 (7.6)
P_10	0.053	0.144 (0.031)	10.9 (5.7)
P_20	0.026	0.106 (0.014)	20.1(6.5)
P_30	0.015	0.091 (0.007)	30.2(7.0)
P_40	0.005	0.085 (0.004)	39.6(7.5)
P_50	0.000	0.083 (0.002)	50.0(7.7)
P_60	-0.004	0.085 (0.004)	60.1(7.7)
P_70	-0.015	0.091 (0.007)	70.2(6.9)
P_80	-0.027	0.106 (0.013)	79.7(6.4)
P_90	-0.056	0.143 (0.031)	88.9(6.0)
Random	-0.003	0.145 (0.015)	50.4(16.0)

We see here the same results as reported before, except that the desired percentages correct are now in line with the percentages that are achieved.

Table 10. 2pl large bank; selection shifted ability level; ; equivalence with random test of 30 items

Selection method	Number of items
Max info	11
p-50	11
p-60	12
p-70	13
p-80	17
p-90	29

Discussion

In this study, it was shown that, in CATs, it is possible to select items with a higher or lower success probability. The selection methods based on the minimal distance between the current ability estimate and the p-points of the items only works satisfactorily if the item bank is calibrated with the 1pl model. This selection method yields bad results when it is applied on an item bank which is calibrated with the 2pl model.

The method introduced, in which items are chosen that have maximum information at an ability lower or higher than the current ability estimate, also performs well in item banks calibrated with the 2pl model. With item banks of a practical size (300), a little loss in measurement precision is the price of a (somewhat) easier (or harder) test. The method is also effective if the selection is combined with the application of exposure control. Getting very high or very low percentages correct was seen to be possible with a larger item bank. In that case, in principle, any desired percentage correct could be reached, but extreme values of the success probabilities are combined with a considerable loss in precision. For practical purposes, selecting, aiming at percentages correct of 60 or 70 (or 40 or 30), seems to be possible without a large loss in precision.

Knowing the effect of selecting with other success probabilities in mind, one could, for CAT applications, build item banks which are more suitable for that purpose. The item banks studied here are in a sense optimal for a CAT with maximum information item selection: the mean difficulty of the items is equal to the mean of the population. If one knows, for instance, that one wants an easy CAT, one could try to construct a bank which is, on average, easier for the population.

Finally, it can be mentioned that all the selection methods and the results are symmetric around the p-50 points. For the selection methods, this is only true for the 1pl and 2pl model. Knowing that the symmetry disappears, it is worthwhile investigating the application of the selection method if the 3pl model, including a guessing parameter, is used.

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.

Eggen, T.J.H.M. (2001). *Overexposure and underexposure of items in computerized adaptive testing.* Measurement and Research Department Reports, 2001-1. Arnhem: Cito.

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

Hambleton, R.K., & Swaminathan, H. (1985) *Item response theory.: Principles and applications.* Boston: Kluwer Academic Publishers

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.

Pitkin, A. K., & Vispoel, W.P.(2001) Differences between self-adapted and computerized adaptive tests: A meta-analysis. *Journal of Educational Measurement, 38*, 235-247.

Rocklin, T.R., & O' Donnell, A.M., (1987). Self-adapted testing: A performance improving variant of computerized adaptive testing. *Journal of Educational Psychology, 79*, 315-319.

Van der Linden, W.J., & Pashley, P.J. (2000). Item selection and ability estimation in adaptive testing. (pp. 1-25) In: W.J. vand Linden, & C.A.W. Glas. (Eds.). *Computerized adaptive testing. Theory and practice.* Dordrecht: Kluwer Academic Publishers

Wainer, H., Dorans, N.J., Flaughter, R., Green, B.F., Mislevy, R.J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A Primer.* Hillsdale (NJ): Lawrence Erlbaum.

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