Psychological Test and Assessment Modeling, Volume 60, 2018 (3), 289-308

An investigation into the usefulness of time-efficient item selection in computerized adaptive testing

Birk Diedenhofen[1] *& Jochen Musch*[2]

Abstract

In computerized adaptive testing (CAT), item-selection algorithms generally attempt to maximize the information provided by each item. However, response times are usually ignored. To improve time efficiency, we established new time-efficient item-selection algorithms that maximize the information collected in a given amount of time. Simulations with 2PL data from the Amsterdam Chess Test (van der Maas & Wagenmakers, 2005) showed that time-efficient item-selection algorithms are indeed able to collect more information in the same amount of time. However, the gains in the amount of information turned out to be rather modest and came at the cost of an increase in measurement bias. For testing practice, our results suggest that item selection based on maximum information can and should be retained as the gold standard in CAT.

Keywords: computerized adaptive testing, item selection, time efficiency, response time, item response theory

[1] *Correspondence concerning this article should be addressed to:* Department of Experimental Psychology, University of Duesseldorf, Universitaetsstrasse 1, Building 23.03, 40225 Duesseldorf, Germany; email: birk.diedenhofen@uni-duesseldorf.de

[2] Department of Experimental Psychology, University of Duesseldorf, Germany

Recent advancements in computer technology and psychometric testing have steadily facilitated the application of computerized adaptive testing (CAT). Unlike linear tests that always present the same fixed set of items to all participants, adaptive tests present tailored sets of items to each participant to measure his or her latent trait or ability level. To this end, item sets are assembled in real time in a computerized environment. An item-selection algorithm chooses items consecutively until a defined stopping criterion is met. Based on a continuously updated estimate of the test taker's ability, item-selection algorithms decide which item will be the best one to present next. Typically, the goal is to estimate a participant's ability with a desired precision using as few items as possible, or as accurately as possible within a given number of items. The superiority of such adaptive testing procedures over conventional tests with regard to precision and efficiency has been widely documented in both simulations and practice (Eggen & Straetmans, 2000; Hornke, 1999).

Item selection and response times

Constructing a tailored test requires a rule for item selection. The most popular method of selecting items is the maximum information algorithm (MI; Barrada, Olea, Ponsoda, & Abad, 2009). This algorithm always selects the item providing the largest amount of information possible at a given ability estimate. When presenting a constant number of items, any deviation from MI by definition leads to a loss of information and thus, poorer estimation precision. Alternatives to MI exist, however. In particular, algorithms that use response times as an additional source of information in the item-selection process have been suggested. In the following, we discuss algorithms that (a) consider response times to reduce the total testing time (Häusler, 2006), (b) try to avoid the drawbacks of speeded testing (Nährer, 1989; van der Linden, Scrams, & Schnipke, 1999), or (c) try to improve the precision of the ability estimate (van der Linden, 2008). We then present new item-selection algorithms that consider response times in addition to the amount of information gained, and report the results of simulation studies that were designed to test whether such item-selection algorithms can beat the maximum information algorithm (MI), the current gold standard for adaptive testing.

Häusler (2006) proposed an algorithm that is based on the observation that employing MI tends to increase item response times because the accompanying constant success probability of $P_i(\theta) = .5$ has negative psychological effects. Participants often perceive tests employing MI as rather challenging (Tonidandel, Quiñones, & Adams, 2002) because participants are better acquainted with the gradual increase in item difficulty that is typical of most tests (e.g., exams given at schools and universities). Consequently, in CAT, participants who are used to conventional testing tend to require more time to feel confident about their answer. To reduce testing time, Häusler (2006) therefore suggested an algorithm that considers the respondent's working style: Respondents who need more time to feel confident about their answers are given items with higher probabilities of success for which shorter response times are to be expected.

A second group of item-selection algorithms makes use of response times to control the speededness of a test. When a time limit is set, participants tend to apply different response strategies to make their individual trade-off between accuracy and speed. Some but not necessarily all participants decide to emphasize speed over accuracy to meet the time limit. As a consequence, the ability measurement is confounded with individual response strategies. To account for diverse response behavior, some item-selection algorithms therefore aim to avoid time limits while still controlling the total test time. In this vein, Nährer (1989) compared *speed-adapted testing* with conventional ability-adapted testing. For each participant, he estimated a linear speed-accuracy trade-off based on an exploratory test phase containing eight items. This individual speed-accuracy trade-off was then used to select eight further items that would most likely not exceed a certain limit placed on the total testing time. Van der Linden et al. (1999) employed a similar procedure that did not rely on an exploratory test phase, but monitored the participants' response times throughout the entire test. To choose an item, their algorithm first selects a complete test that meets a predefined constraint on the total test time and provides maximum information given the current ability estimate. Finally, the item that provides the greatest amount of information out of the selected items is administered.

A common empirical finding is that response time is related to ability such that more able participants typically answer more quickly. Various models have been proposed to incorporate response times into ability estimation. Schnipke and Scrams (2002) provide an overview of some of the most popular approaches. For example, Thissen (1983) suggested an item response time model based on the usually positive correlation between ability and speed. The model includes slowness parameters for both participants and items. Pursuing a similar approach, van der Linden (2007) established a hierarchical framework for modeling speed and accuracy. His approach allowed for the combination of different models of responses and response time distributions on a first level, and different models of the distributions of ability and speed parameters on a second level. In an application of this framework, van der Linden (2008) was able to substantially improve the accuracy of the ability estimates by taking response times into account when selecting the next item for presentation. His procedure determined the current ability estimate after the presentation of each item, not only from the participant's responses but also from the corresponding response times. The usual maximum information algorithm (MI) was then used to select the next item. In a recent publication, Hohensinn and Kubinger (2017) investigated whether speed and power unidimensionally measure the same ability by comparing different modeling approaches. They found that a multi-dimensional polytomous Rasch model employing a joint measurement approach was more appropriate than both, a speed-and-power two-steps model assuming speed and power to be completely independent and a unidimensional Rasch model assuming speed and power to represent the same ability.

The present study

Comparisons between adaptive tests using MI and fixed-item tests showed savings in the number of items ranging from 22 % (Eggen & Straetmans, 2000) up to 50 % (Hornke, 1999) for the adaptive tests. Thus, a considerably smaller number of items are necessary to achieve the same level of precision when a test is conducted adaptively. Wild (1989) also found item savings, but surprisingly, there was almost no difference in total test time between adaptive and conventional testing in her study. She therefore argued that the economic benefits of CAT may actually vanish when response times are taken into account. This finding raises the question of whether MI is indeed the universally best item-selection procedure. In Wild's (1989) study, high response times for a subset of items nullified the item savings and, thus, made the economical benefits of CAT disappear. Her findings revealed a systematic failure from which the present practice of CAT suffers: For decades, item-selection algorithms have typically been judged only by the amount of information gained after the presentation of a given number of items. Response latencies were either disregarded or simply assumed to be equal for each item. However, tests that are based on the same number of items will rarely result in identical total testing times. Completely disregarding response times may be considered particularly problematic with regard to power tests, which do not force the respondents to stay within a given time limit. Presumably, power tests would allow test administrators to better distinguish between different levels of ability once response times are taken into account.

In the present paper, we suggest that the efficiency of a test be assessed by determining the estimation precision achieved in a given amount of time. Accordingly, we implemented a family of algorithms that maximize the amount of information gained per second, instead of maximizing the amount of information gained per item. The rationale of all algorithms is that a larger number of less informative items that are answered more quickly may provide more information in the same amount of time than a small number of items that provide maximum information but that consume disproportionate amounts of time. We investigated whether such algorithms would be able to provide a higher precision of measurement without extending the testing time or provide shorter testing times without sacrificing accuracy.

First, we established a maximum information per second algorithm (MIPS), which uses the ratio of item information to response time as a criterion for item selection. In item response theory, the information an item provides at a given ability level can easily be calculated using the information function. Like MI, to select the item that is to be presented next, MIPS determines the amount of information provided by each of the items that is still left in the pool. However, the time a participant will need to answer one of these remaining items is unknown. A time-efficient item-selection algorithm must therefore be able to predict a participant's response time for each item. In this study, we therefore implemented three variants of MIPS that differ in how response times are estimated based on data from previous administrations of the test. The first variant of MIPS—MIPS-omniscient—simply uses the original response times and thus, perfectly

predicts the participants' response times in our simulation. It served as a baseline for scrutinizing the improvements that would be theoretically possible through time-efficient item selection for a given set of test data. A related approach—MIPS-average—relies on the mean response time to an item across all participants. A third variant of MIPS—MIPS-regression—uses the correlation between ability and response time in a regression model that predicts the time a participant needs to respond to an item. All variants of MIPS divide the expected information gain by the predicted response time to derive an estimate of the information gain per second provided by an item. Each MIPS variant eventually administers the item that contributes the most information per second. Thus, our approach is based on the expected response times for items still left in the pool, whereas van der Linden's (2008) approach considers only response times for items that have already been administered and makes no predictions regarding the response times of future items. Moreover, van der Linden's (2008) approach considers response times only in an indirect way. Response times are used to update the current ability estimate, but the selection of the next item is based on maximizing information without regard to the expected response time.

The aim of the present study was to investigate whether algorithms that take response times into account by following the MIPS principle will outperform MI with regard to how much information can be accumulated in a given amount of time. By definition, MI is the most informative algorithm on an item basis and is therefore the current gold standard in CAT. New algorithms have to stand up to a comparison with MI. To this end, we let the algorithms compete against each other in simulations based on empirical test data. To create conditions that permitted us to scrutinize differences in testing time, items from a domain with a known high variability in response latencies were chosen for the simulations. We therefore used cognitively demanding chess problems characterized by large and highly variable response latencies. In our simulations, several indices were calculated after the administration of each item. First, we computed the standard error and the estimation bias to compare the algorithms with regard to the precision of their ability estimation. Because the true abilities of the participants were unknown, we operationalized the bias as the difference between the current ability estimate and the best possible ability estimate based on the full set of all items. We also determined the total test time to explore the time efficiency of the competing algorithms. To compute an index of similarity of the competing algorithms—termed *MI overlap*—we calculated the percentage of items that were chosen not only by the respective algorithm but also by MI. This index allowed us to quantify the differences in the item sets created by employing different item-selection algorithms.

By definition, selecting items according to MI gathers the most information per item. Whether MIPS can outperform MI by collecting more information per second than MI should depend on the quality of the MIPS response-time predictions. Apart from that, the magnitude of the difference in performance between MIPS and MI should be influenced by the variance in the items' response times. A potential advantage of MIPS over MI should be large when items differ widely with regard to their response times. A larger

pool of items is more likely to entail a larger variability in response times and thus, is more likely to result in the choice of different sets of items when different item-selection rules are employed. We also expected the potential advantage of MIPS over MI to be large at the beginning of the test when MIPS has the greatest freedom to present items that not only collect a lot of information, but that can also be answered more quickly than the remaining items. At later stages of the test, this freedom should be gradually diminished because fewer items with desirable characteristics are left in the item pool.

A potential superiority of MIPS over MI would also show itself in a small MI overlap, i.e., the set of intersecting items selected by both MIPS and MI. The more overlap there is between the items that are chosen by the two selection rules, the less room there is for a potential superiority of MIPS over MI because obviously, MI cannot be outperformed by an algorithm that chooses the same items. In our simulations, we also addressed a potential drawback of selecting items based on time efficiency. We investigated whether employing MIPS would result in a larger estimation bias because a preference for selecting items with short response times may necessarily be accompanied by a tendency toward the selection of easier items. Given that test bias is calculated as the difference between the current ability estimate and the ability estimate based on all items, ability estimates that are mainly based on easier items may potentially be affected by overestimation bias.

Method

Amsterdam Chess Test

For the simulations, we used previously recorded responses to the choose-a-move tasks A and B of the Amsterdam Chess Test (ACT; van der Maas & Wagenmakers, 2005). The ACT data used for our simulations were collected by van der Maas and Wagenmakers (2005) in a large sample of chess players during an Open Dutch Championship in Dieren. The two subtests consist of 40 chess problems each for which the participants were asked to find the best possible move from a given position as quickly as possible. Figure 2 shows two items of the choose-a-move task. The chess positions were presented on a computer screen, and participants entered their moves using the computer mouse. Participants were asked to answer each item within 30 s. If no correct solution was entered within the time limit, the item was scored as wrong. To facilitate the subsequent simulations and because there was only a very small percentage of missing data, only 249 complete cases were used for analysis. Another 10 participants were excluded from the analysis because they failed to provide an answer to one or more of the choose-a-move items.

Item parameters for two dichotomous latent trait models—the 1PL model and the 2PL model—were estimated from the participants' response patterns using Bilog-MG 3 (Zimowski, Muraki, Mislevy, & Bock, 2003). The 2PL model yielded a significantly better fit ($\chi^2(80) = 571.59$, $p < .001$) and was therefore used exclusively for the following

analysis, with the exception of four items that deviated from the model according to χ^2 item-fit tests provided by Bilog (all $p < .05$). The four items that mismatched the 2PL model were the two easiest (items A31 and B31) and the two most difficult items (items B30 and B40) in the pool. Excluding these four items left 76 items that were used in the simulations.

Table 1:
Distribution of the participants' ability $\hat{\theta}$ and the discriminatory power, difficulty, and response time for the 76 items.

Parameter	***M***	***SD***	**min**	**max**
Participants' ability $\hat{\theta}$	-0.06	1.16	-3.06	3.00
Item discriminatory power a	0.86	0.35	0.30	1.73
Item difficulty b	0.02	1.77	-4.53	4.98
Item response time T (s)	13.36	5.13	3.34	22.90

The distribution of discriminatory power, difficulty, and response time for the 76 items are displayed in Table 1. Figure 1 displays the density plot of the mean item response time. Response time was correlated with ability: Participants achieving a higher test score answered the items more quickly on average ($r = -.43$, $p < .001$).

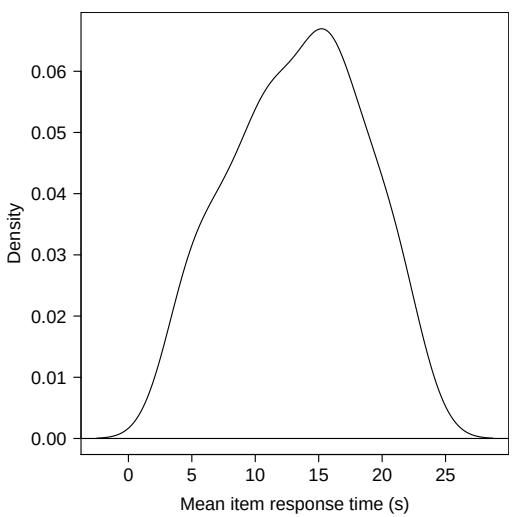

Figure 1:
Density plot of mean item response times.

Simulations

The simulations were carried out using R (R Core Team, 2017a). We used the *foreign* package (R Core Team, 2017b) to import data into R. A participant's ability $\hat{\theta}$ was estimated using the expected a posteriori (EAP) algorithm (Bock & Mislevy, 1982). The distribution of the participants' ability $\hat{\theta}$ is displayed in Table 1. In the present study, we assumed the prior distribution to be standard normal and used $K = 100$ quadrature points to achieve a high level of estimation precision. The points and weights were obtained from the Gauss-Hermite quadrature using the *statmod* package (Smyth, 1998).

The item pool used in the simulations consisted of the 76 items from the ACT detailed above. From this pool, the item-selection algorithms—described in more detail in the following section—chose items successively. To simulate a computerized adaptive test for each of the 249 participants, we used the respective test taker's responses that were recorded when he or she completed the entire item set of the ACT (van der Maas & Wagenmakers, 2005). Thus, based on real data collected on a complete test, we were able to simulate the results of hypothetical computerized adaptive tests based on different item-selection rules. To this end, only the responses to a subset of all items—the subset of items that was selected by the respective item-selection rule—had to be taken into account. Due to the diverse item-selection rules, the competing algorithms produced tests with different item orders that therefore exhibited different psychometric properties. After each item was presented, the test properties of the items that had been selected up to that point were determined for comparison. In the simulations, each participant began the adaptive test with an initial ability estimate of $\theta = 0$. Each item cycle was carried out as follows: A given item-selection rule was applied to choose the next item, which was then answered according to the participant's original response pattern from the ACT. Next, the information gained from the new item was used to update the estimate of the participant's ability. This procedure was repeated until all 76 items had been administered. If the selection criterion of any algorithm yielded ambiguous results—two or more items met the selection criterion equally well—, the next item was drawn randomly from the items closest to the criterion.

After the presentation of each item, the following indices were computed as dependent variables separately for each item-selection algorithm. First, we determined the standard error as the square root of the reciprocal Fisher information and the bias of the estimation by subtracting the current ability estimate from the ability estimation based on the complete set of items. To scrutinize time efficiency, we determined the total test time of all participants by summing their response times using the response times recorded during the original ACT administration. The percentage of items selected by both a given algorithm and MI was computed to form an *MI overlap* index. All of these indices were determined on an item level, i.e., every index was updated after the presentation of each item. However, a common time scale was required to compare item-selection algorithms with respect to their time efficiency. To this end, rather than as a function of the number of items presented, all indices were plotted against the total test time.

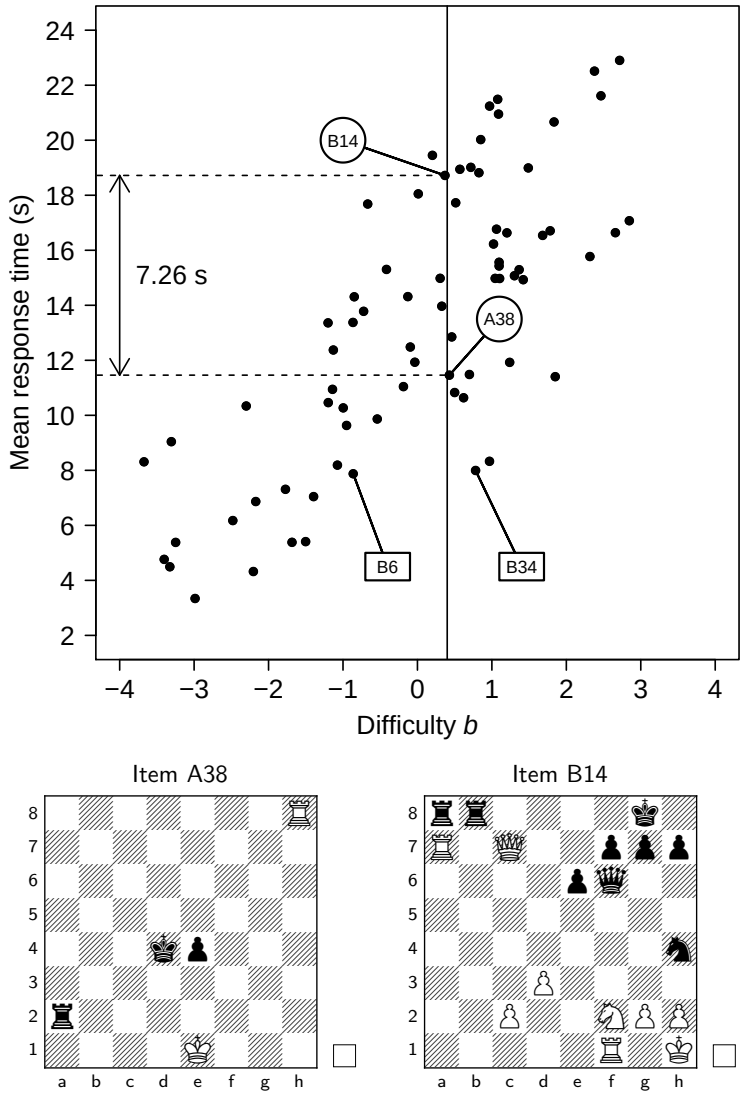

Figure 2:

For all items, their mean response time across all participants is plotted against their difficulty. For a participant with an ability of 0.4 (black vertical line), item B14 does not provide significantly more information than item A38, but these two items differ vastly in their mean response times. The maximum information algorithm (MI) would select item B14 because it provides slightly more information, whereas the maximum information per second algorithm (MIPS-average) would choose the more time-efficient item A38. This would allow MIPS-average to administer another item (e.g., B6 or B34) and thus to collect additional information using the time saved. The two chess diagrams display items A38 and B14 of the Amsterdam Chess Test. The best moves in these positions are 1. … Rh3 (Philidor defense) and 1. Qb7 (overloading), respectively.

Item-selection algorithms

In the following, we describe the six item-selection algorithms that we employed in the simulations.

Maximum information algorithm (MI). As stated above, MI determines the amount of information provided by an item given the current ability estimate and selects the item that contributes the most information. The trade-off between proximity to the ability estimate and discriminatory power is accomplished by maximizing $I_i(\theta)$.

MIPS-omniscient. Like MI, MIPS-omniscient calculates the potential information gain $I_i(\theta)$ at the present ability estimate. To maximize time efficiency, the algorithm needs information about the expected time a participant will take to answer an item. Therefore, the participant's response time T_i for an item i is additionally taken into account. The algorithm eventually selects the item providing the largest amount of information per second: $max(I_i(\theta)/T_i)$. MIPS-omniscient is an algorithm that cannot be applied in real testing situations because the time a participant needs to answer an item is usually unknown in advance. In the following simulations, however, this missing information was taken from the previous ACT administration in which every item of the complete test had already been presented to each participant. MIPS-omniscient was thus made omniscient by having access to these previously recorded response times prior to the administration of these item. Thus, MIPS-omniscient quantifies the maximum improvement that can theoretically be achieved using a time-efficient item-selection rule. Both MI and MIPS-omniscient served as baselines for comparison with the remaining algorithms, along with a random selection algorithm (RANDOM) and an algorithm that always picks the item with the fastest average response time regardless of the amount of information it provides (FASTEST).

Two additional variants of item-selection algorithms that followed the principle of maximizing information per second were also employed. Unlike MIPS-omniscient, these algorithms can also be applied in situations in which the participant's presumed response time is unknown at the time of testing because the algorithms predict the time T_i a participant takes to answer a given item based on the original ACT data. The algorithms differed only with respect to how they predicted T_i, i.e., with regard to how they used the original ACT data to provide an estimated response time. Time efficiency can be optimized using these algorithms only if the algorithms' predictions get as close as possible to the participant's true response times. Thus, the time efficiency of these item-selection algorithms can only be as good as their prediction of the participants' response times.

MIPS-average. MIPS-average uses the average response time of all participants excluding the participant currently being tested to predict the response time T_i for an item, with the aim of maximizing $I_i(\theta)/T_i$. The potential advantage of MIPS-average over MI can be demonstrated by the following example: For a participant with θ = 0.4, item A38 (a = 1.03, b = 0.43) and item B14 (a = 1.03, b = 0.37; see Figure 2) offer nearly the same amount of information (0.264 and 0.265, respectively) in that they have equal discriminatory power, and their difficulties are equally distant from the current ability estimate.

Item B14 can only be solved after a tactical calculation involving all possible reactions to a queen sacrifice that is difficult to spot and that has to be identified from a large number of candidate moves, whereas item A38 essentially tests whether or not a participant knows the so-called "Philidor defense" in rook and pawn versus rook endgames (de la Villa, 2015). For strong players, this knowledge is readily available from long-term memory and therefore does not require extensive calculations. From among these two items, MI would select item B14 because it provides slightly more information than item A38. However, with a mean response time of 11.46 s, item A38 is answered much faster than item B14, which takes 18.72 s to complete on average. According to these average response times, item A38 gathers almost twice as much information per second as item B14. Unlike MI, MIPS-average takes advantage of this discrepancy and decides in favor of item A38 in spite of the slightly inferior expected gain in information. The expected time savings of 7.26 s allows MIPS-average to administer an additional item (e.g., item B6 or B34). In this example, MIPS-average therefore provides considerably more information than MI within the same period of time.

MIPS-regression. An item's response time often depends on the participant's ability: The higher the ability level, the faster the participant's response. For the ACT, both a linear regression, $T_i = b_0 + b_1 x + \epsilon$ ($b_1 = -1.28$, $t(247) = -7.62$, $p < .001$) and a quadratic regression, $T_i = b_0 + b_1 x + b_2 x^2 + \epsilon$, ($b_1 = -1.35$, $t(246) = -8.00$, $p < .001$; $b_2 = -.29$, $t(246) = -2.48$, $p < .05$), showed that the participants' abilities predicted their mean response times. The linear regression ($R^2 = .19$, $F(1, 247) = 58.10$, $p < .001$) as well as the quadratic regression ($R^2 = .20$, $F(2, 246) = 32.74$, $p < .001$) explained a significant proportion of variance in mean response times (R^2 was adjusted for the number of variables in the regression according to Faraway, 2004, p. 136). In a direct comparison, the quadratic regression turned out to fit the data significantly better ($F(1, 246) = 6.18$, $p < .05$). An analysis at the item level confirmed that the relation between ability and response time could be modeled using a quadratic regression. For 65 of the 76 items, the quadratic regression was significant. To maximize the collected information per second ($max(I_i(\theta)/T_i)$), MIPS-regression therefore used quadratic regressions to predict the participants' response times T_i. In the simulated adaptive test, the response time for each item was predicted using the data of all participants excluding the participant currently being tested.

Results

In Figure 3, the mean number of items that were presented by the algorithms during the first 500 s of the test is displayed. As expected, the algorithm that always selected the fastest item (FASTEST) administered the most items in the first 500 s, followed by MIPS-omniscient, MIPS-regression, MIPS-average, and RANDOM. Of all algorithms, MI administered the least number of items during the first 500 s of the test. Figure 4 shows the mean cumulative testing time for the competing algorithms as a function of the number of items presented and confirms this pattern; whereas no algorithm needs

less time to present the first 15 items than FASTEST, MI needs more time for the first 15 items than any of the more time-efficient MIPS alternatives.

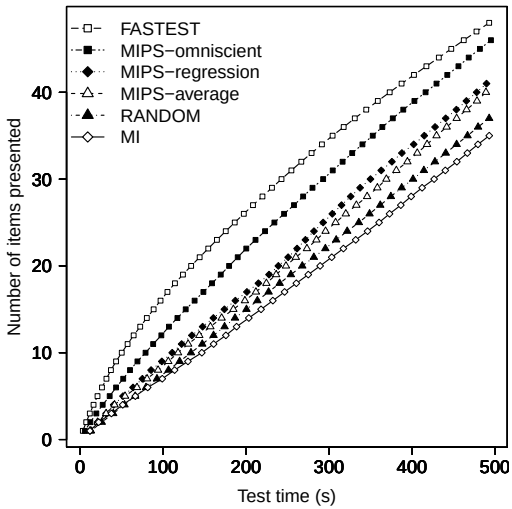

Figure 3:
Mean number of items presented by the algorithms during the first 500 s of the test.

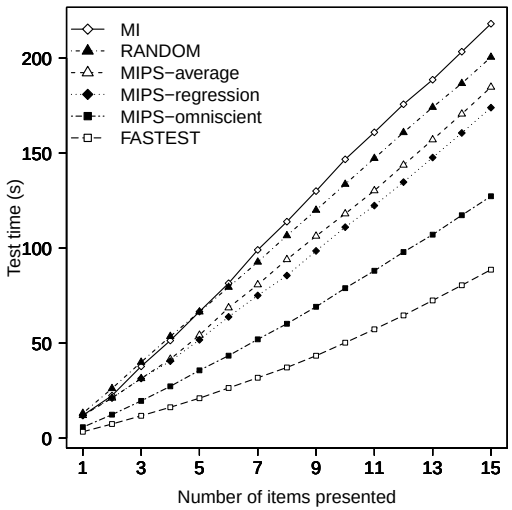

Figure 4:
Mean cumulative testing time for the competing algorithms as a function of the number of items presented.

In general, all algorithms exhibited smaller standard errors when more items had already been selected. As expected, a random item selection (RANDOM) led to the largest standard error at the beginning of the test, followed by FASTEST (Figure 5). MI yielded considerably lower standard errors than RANDOM and FASTEST. However, all algorithms following the MIPS principle were able to reduce uncertainty even faster than MI. The advantage of the MIPS algorithms was greatest at the beginning of the test but declined toward the end because the steadily decreasing number of items that were left available for selection reduced item choice and thus enforced a convergence of all algorithms toward the end of the test. As can be seen in Figure 5, MIPS-regression turned out to be even slightly more time-efficient than MIPS-average. Beyond a testing time of 200 s, it was no longer possible to differentiate between MIPS-average and MIPS-regression, however. As expected, MIPS-omniscient provided the fastest decrease in standard error. The standard error of all algorithms converged at 0.305 at a total test duration of 1015.61 s after the completion of the entire test consisting of 76 items.

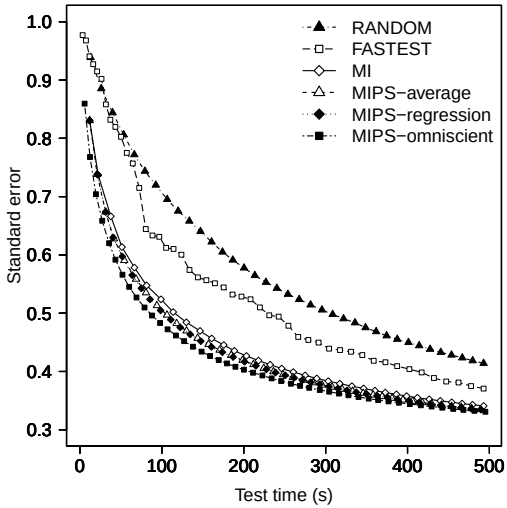

Figure 5:
The standard error for item-selection algorithms during the first 200 s of testing time. All algorithms of the maximum information per second algorithm (MIPS) family reduce measurement error more quickly than the maximum information algorithm (MI).

A comparison of the total testing time needed to achieve a desired level of precision revealed the time savings that could be obtained by using different item-selection algorithms. A maximum reduction in the total testing time of 73.40 s (7.2 %) was achieved by using MIPS-omniscient instead of MI. For MIPS-regression, the maximum time savings was 48.49 s (4.8 %).

To compute an index of the magnitude of the improvement due to the use of time-efficient item-selection procedures over MI, we divided the standard error associated with each item-selection algorithm by the standard error associated with the use of MI. As Figure 6 shows, at the beginning of the test in particular, MIPS-omniscient was able to outperform MI by about 8 %; using MIPS-omniscient provided a standard error that amounted to only 92 % of the standard error associated with MI. For MIPS-regression and MIPS-average, respectively, the maximum reduction in the standard error relative to MI was 4 %. For each algorithm, Table 2 shows the standard error relative to MI after averaging across all items.

Table 2:

The standard error of the ability estimates for different algorithms, with MI as the reference algorithm. Values below 1 indicate a reduced standard error, and values above 1 an increased standard error as compared to MI.

Algorithm	Average standard error relative to MI
MIPS-omniscient	0.961
MIPS-regression	0.982
MIPS-average	0.983
FASTEST	1.130
RANDOM	1.198

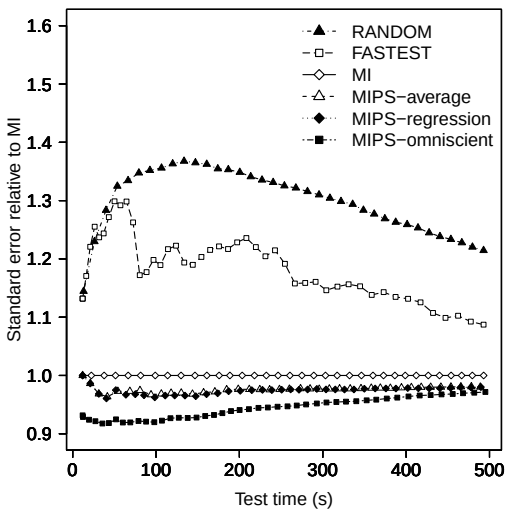

Figure 6:

The standard error relative to the maximum information algorithm (MI) during the first 500 s of the test.

For many algorithms, bias increased during the first 100 s of the test (Figure 7). FASTEST reached the largest bias of 0.43 after 65 s because its propensity to always select the fastest item led to a particularly strong preference for the easiest items at the beginning of the test. However, even the bias of FASTEST gradually declined toward zero at the end of the test. MIPS-omniscient also had a notably large bias. Less bias was shown by MIPS-regression and MIPS-average. The ability estimates by MI and RANDOM were least biased.

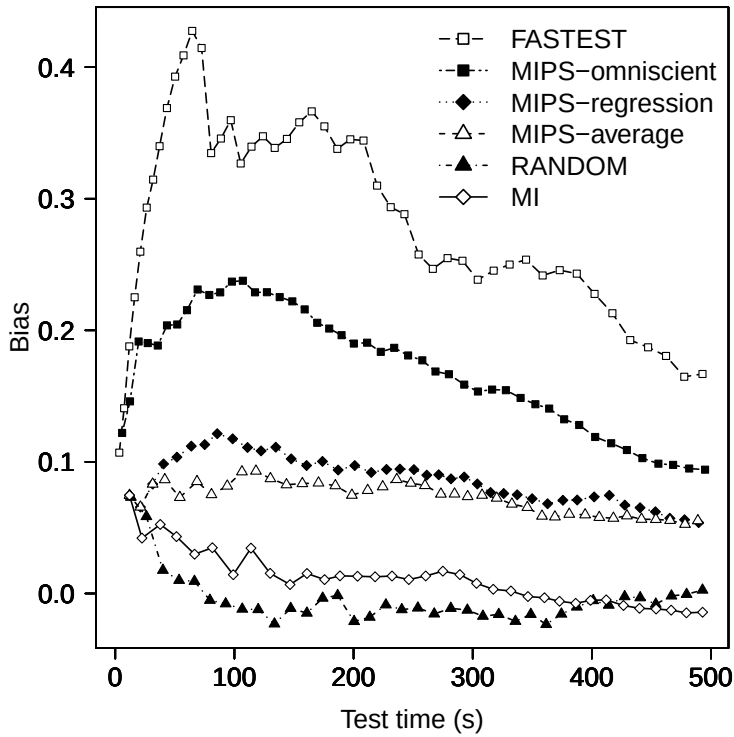

Figure 7:
Estimation bias for all algorithms during the first 500 s of the test. The final ability estimate based on the complete set of 76 items was used as a reference. Therefore, the bias of all algorithms approaches zero toward the end of the test when all items had been presented.

The MI overlap index was computed as the percentage of items selected by a given algorithm that were also selected by MI (Figure 8). FASTEST and RANDOM showed the smallest MI overlap. Due to the exhaustion of the item pool toward the end of the test, MI overlap gradually increased when items were selected randomly or based exclusively on their speed. MI overlap was larger for item-selection algorithms that considered both information gain and response time.

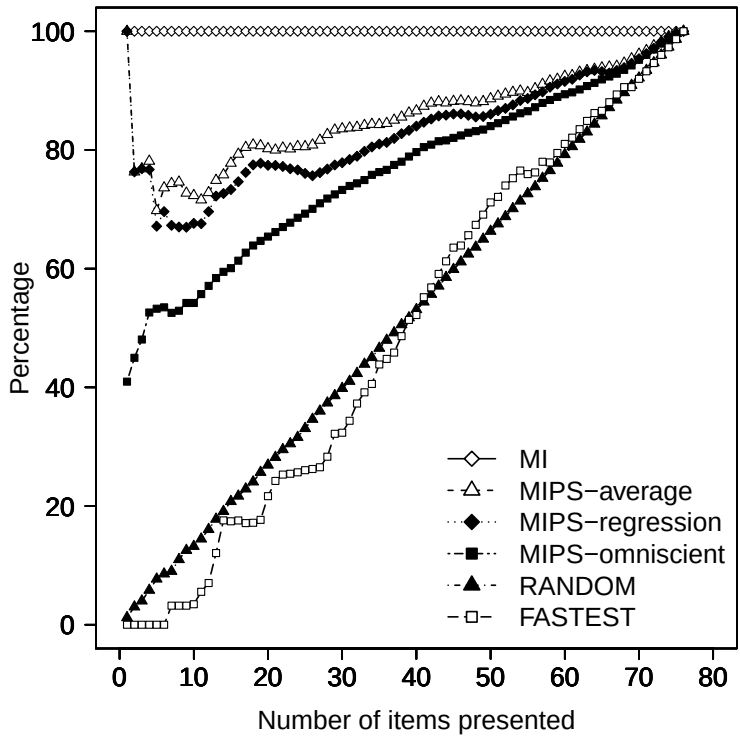

Figure 8:
MI overlap plotted against the number of items presented. MI overlap is the percentage of items selected by both a given algorithm and the maximum information algorithm (MI).

Discussion

The aim of the present study was to investigate whether time-efficient item-selection algorithms could outperform MI with regard to the estimation precision achieved in a given amount of time. Unlike MI, which maximizes the information gathered per item, MIPS algorithms maximize the information gained per second. The results of our computer simulations showed that item-selection algorithms based on the MIPS principle were indeed able to outperform MI in terms of time efficiency. Standard error reductions were greatest at the beginning of the test, whereas the largest time savings were achieved toward the end of the test. Overall, the effects turned out to be relatively small, however. MIPS-omniscient was most time-efficient with a maximum reduction of 0.06 compared to MI. Relative to MI, MIPS-omniscient provided a reduction in the standard error of about 8 % and a maximum time savings of up to 73.40 s (7.2 %).

It is important to remember that MIPS-omniscient can only be employed in simulation environments in which the participants' response times are known in advance. However, this algorithm provides useful information about the degree to which time efficiency can optimally be improved for a given set of items. Regarding the two algorithms that predicted future response times, MIPS-average yielded less information per second than MIPS-regression. MIPS-regression decreased the standard error by 0.026 relative to MI and achieved a reduction in the standard error of 4 %. MIPS-regression also provided a time savings of up to 48.49 s (4.8 %). It is important to note, however, that these figures represent the maximum possible improvements that were obtained. Under less optimal circumstances, the improvements turned out to be smaller; however, no MIPS algorithm ever fell below the efficiency of MI. The results of FASTEST demonstrate that a strategy that is focused solely on time savings will not be successful in accurately assessing ability in a time-efficient manner.

The comparison of MIPS-omniscient and MIPS-regression indicates that there is still room for a further improvement of time-efficient item-selection algorithms. To tap the full potential of MIPS, more accurate predictions of the participants' response times are necessary. This can be achieved by more advanced response-time estimation methods. For example, the participants' response times for previously answered items could additionally be used to improve response-time predictions (van der Linden, 2008).

As the MI overlap index indicated, there were a sufficient number of items with an advantageous information-to-time ratio in the pool to allow time-efficient algorithms to outperform MI by selecting different items. In general, with an increasing number of available items, a smaller MI overlap and a larger potential gain from employing time-efficient algorithms may be expected.

Some limitations of the present study must be acknowledged. The generalizability of the results might be limited to tests that are similar to the ACT, i.e., complex cognitive tasks with highly variable response times. Furthermore, we used the same data for both item calibration and the subsequent simulations. In a real-world application, the populations for calibration and testing will likely deviate, and as a consequence, the precision of response-time predictions may deteriorate. Considering that the conditions of our simulations were rather favorable for time-efficient item-selection algorithms, even the small improvements found in our simulations might be difficult to achieve in real-world settings.

On the other hand, the ACT employed a time limit of 30 s per item. Possibly, a pure power test would have created more favorable conditions for the application of algorithms that consider response times in addition to solution rates. However, it is important to note that in spite of the time limit that was used, response times in the ACT showed considerable variation (SD = 8.52 s), and the time limit ended responses in only 5.6 % of the trials. The present ACT items therefore met the main requirements for a successful application of item selection algorithms that consider response times in addition to solution rates. However, future investigations should scrutinize whether items presented

without any time limit lead to even larger time savings that make it more worthwhile to apply time-efficient item selection algorithms.

In our simulations, we were not able to investigate possible motivational effects associated with the use of time-efficient testing. However, there is reason to expect that employing MIPS algorithms may have some motivational benefits. This is because item-selection algorithms that prefer shorter response times also tend to select easier items, and this helps to improve the participant's self-confidence (Häusler & Sommer, 2008).

A major drawback associated with the use of time-efficient item-selection algorithms is an increase in bias. Although MIPS algorithms were able to decrease standard errors when compared to MI, they also introduced considerably more bias. Thus, the increase in reliability that can be achieved by using time-efficient item-selection algorithms comes at the cost of a decrease in validity. This is because a selection of easy items from the item pool results in higher ability estimates on average than estimations based on the entire item pool. To further investigate the generalizability of our findings, it would be interesting to scrutinize additional empirical data sets. This would also allow us to determine which item response time distributions best facilitate time-efficient testing.

To summarize, we were able to show that under favorable conditions, time-efficient item-selection algorithms are able to outperform MI. Even under such favorable conditions, however, the achieved increase in precision was rather modest. Moreover, the reduction in standard errors by considering response times came at the cost of an increase in estimation bias. For this reason, the results of the present study do not support the suspicion that MI might be a suboptimal item-selection algorithm because of its lack of consideration of response times. Until further notice, MI should be retained as the gold standard for item selection in computerized adaptive tests.

References

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2009). Item selection rules in computerized adaptive testing: Accuracy and security. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *5*, 7–17. doi: 10.1027/1614-2241.5.1.7

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431–444. doi: 10.1177/014662168200600405

de la Villa, J. (2015). *100 Endgames You Must Know: Vital Lessons for Every Chess Player* (4th ed.). Alkmaar, The Netherlands: New In Chess.

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. doi: 10.1177/00131640021970862

Faraway, J. J. (2004). *Linear models with R*. Chapman and Hall/CRC.

Hohensinn, C., & Kubinger, K. D. (2017). Using Rasch model generalizations for taking testee's

speed, in addition to their power, into account. *Psychological Test and Assessment Modeling*, *59*, 93–108.

Hornke, L. F. (1999). Benefits from computerized adaptive testing as seen in simulation studies. *European Journal of Psychological Assessment*, *15*, 91–98. doi: 10.1027//1015-5759.15.2 .91

Häusler, J. (2006). Adaptive success control in computerized adaptive testing. *Psychology Science*, *48*, 436–450.

Häusler, J., & Sommer, M. (2008). The effect of success probability on test economy and self-confidence in computerized adaptive tests. *Psychology Science*, *50*, 75–87.

Nährer, W. (1989). „Schnelligkeitsangepasstes Testen": Testökonomie unter Berücksichtigung der Testzeiten ["Speed-adapted testing": Test economy with regard to test duration]. In K. D. Kubinger (Ed.), *Moderne Testtheorie [Modern test theory]* (pp. 219–236). Weinheim, Germany: Beltz.

R Core Team. (2017a). *R: A language and environment for statistical computing*. Vienna, Austria. Retrieved from `http://www.R-project.org`

R Core Team. (2017b). *foreign: Read data stored by 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ... [Computer software]*. Retrieved from `https://CRAN.R-project.org/package=foreign` (R package version 0.8-69)

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. Potenza, J. J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237–266). Hillsdale, NJ: Lawrence Erlbaum.

Smyth, G. K. (1998). Numerical integration. In P. Armitage & T. Colton (Eds.), *Encyclopedia of biostatistics* (pp. 3088–3095). London: Wiley.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 179–203). London: Academic Press.

Tonidandel, S., Quiñones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology*, *87*, 320–332. doi: 10.1037//0021-9010.87.2.320

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308. doi: 10.1007/s11336-006-1478-z

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*, 5–20. doi: 10.3102/1076998607302626

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*, 195–210. doi: 10.1177/01466219922031329

van der Maas, H. L. J., & Wagenmakers, E. J. (2005). A psychometric analysis of chess expertise. *American Journal of Psychology*, *118*, 29–60.

Wild, B. (1989). Neue Erkenntnisse zur Effizienz des „tailored"-adaptiven Testens [Recent findings for the efficiency of tailored-adaptive testing]. In K. D. Kubinger (Ed.), *Moderne Testtheorie [Modern test theory]* (pp. 179–186). Weinheim, Germany: Beltz.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3 for Windows: Multiple-group IRT analysis and test maintenance for binary items [Computer software]*. Lincolnwood, IL: Scientific Software International, Inc.