Proceedings of the Federated Conference on Computer Science and Information Systems pp. 359–367
DOI: 10.15439/2020F197
ISSN 2300-5963 ACSIS, Vol. 21

Feasibility of computerized adaptive testing evaluated by Monte-Carlo and post-hoc simulations

Lubomír Štěpánek
Institute of Biophysics and Informatics
First Faculty of Medicine, Charles University
Salmovská 1, Praha 2
lubomir.stepanek@lf1.cuni.cz

Patrícia Martinková
Institute of Computer Science of the Czech Academy of Sciences
Pod Vodárenskou věží 2, Praha 8
Faculty of Education, Charles University, Myslíkova 7, Praha 1
martinkova@cs.cas.cz

***Abstract*—Computerized adaptive testing (CAT) is a modern alternative to classical paper and pencil testing. CAT is based on an automated selection of optimal item corresponding to current estimate of test-taker's ability, which is in contrast to fixed predefined items assigned in linear test. Advantages of CAT include lowered test anxiety and shortened test length, increased precision of estimates of test-takers' abilities, and lowered level of item exposure thus better security. Challenges are high technical demands on the whole test work-flow and need of large item banks.**

In this study, we analyze feasibility and advantages of computerized adaptive testing using a Monte-Carlo simulation and post-hoc analysis based on a real linear admission test administrated at a medical college. We compare various settings of the adaptive test in terms of precision of ability estimates and test length.

We find out that with adaptive item selection, the test length can be reduced to 40 out of 100 items while keeping the precision of ability estimates within the prescribed range and obtaining ability estimates highly correlated to estimates based on complete linear test (Pearson's $\rho \doteq 0.96$). We also demonstrate positive effect of content balancing and item exposure rate control on item composition.

I. Introduction

MULTI-ITEM assessment instruments find their use in number of areas including admission or other educational tests, psychological measurement, health-related questionnaires, and other behavioral measurements. A usual way to perform achievement testing is by assigning a fixed set of items which are supposed to measure construct of interest, such as knowledge of biology, level of depression, fatigue, or respondent's quality of life.

Given that the abilities may greatly differ across test-takers, the respondents with higher levels of ability may be bored by easier items, while those with lower levels of ability might experience inconvenient stress. An effective and appropriate selection of items which suit the best the test-takers of a given ability can thus be more convenient for respondents, may save time and moreover provide estimates of better precision than fixed tests of the same length.

Adaptive tests [1], [2] have been an alternative to linear tests for decades. The most complex version of adaptive tests is the one in which the item selection is done after each item administration depending on the current estimate of test-taker's ability which is iteratively updated. Multistage tests [3] on the other hand involve assigning blocks of items adaptively depending on the ability estimate from the previous test section.

Basic principles of computerized adaptive test are presented in Figure 1.

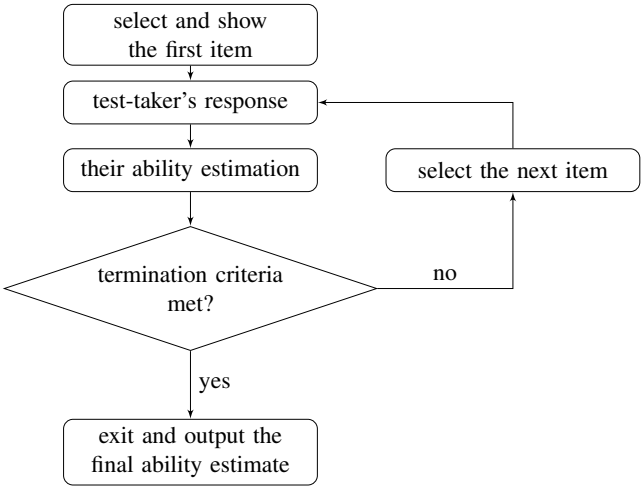

Fig. 1. Computerized adaptive testing flowchart

An adaptive test is initialized by the selection and administration of the first item. The first item can be selected randomly or based on prior ability estimate of the respondent. Average ability can be used as an uninformed estimate, alternatively, initial estimate may be based on respondent's answers to one or a small number of pre-test items.

Depending on the answer to the first item, the test-taker ability estimate is updated. If the termination criterion (such as number of administered items or precision of the estimate) is not met, the updated ability estimate is used to select the next optimal item. This cycle is repeated until the a priori specified termination criterion is met; then, eventually, the test is stopped and final estimate of the test-taker ability is provided as an output.

A. Comparison of linear and adaptive testing

Both the linear and adaptive test scenario have their advantages and disadvantages, respectively. Advantages of adaptive

tests have been demonstrated in areas of educational testing [4], testing of psychological distress [5], [6], as well as health-related measurements such as in mobility surveys [7], and testing general disability [8]. While the adaptive tests are usually shorter in terms of number of items and overall time needed to complete the test, they also enable to estimate test-taker's ability with better precision than linear tests of similar length. The lower level of item exposure usually implies also better security when items are administered adaptively.

However, since the adaptive testing is more complex it requires higher technical facility and support of trained experts. The initial setting of the adaptive test may provide number of options which may have crucial impact on functioning of the adaptive test. Therefore, feasibility and optimal setting of CAT with respect to the given item bank and population of test takers need to be analyzed in order to apply the adaptive test effectively and profitably.

In this work, we use Monte Carlo simulations and post-hoc analysis based on real data of admission test administrated at a medical college with the aim to derive the optimal setting of adaptive test. We also compare the precision of different settings and estimate the correlation between the adaptively estimated ability and estimates based on answers to complete set of 100 items. We discuss results for different levels of precision, and various test termination criteria. We also implement content balancing and item exposure rate control to see how it affects performance and properties of the adaptive test. We discuss the findings in context of the admission testing and other educational testing at medical faculties.

The paper proceeds as follows. We firstly describe the data and introduce all necessary background theory, including underlying models, settings of adaptive tests and design of the simulation studies in the *Research Methodology* section. We then present results of the post-hoc analysis and Monte Carlo simulation in Section *Results*. Finally, discussion and final remarks are provided in *Conclusion* section.

II. Research Methodology

A. Data and item calibration

We used data from a real fixed admission test administrated to 2372 test-takers (applicants) at First Faculty of Medicine, Charles University, Prague in 2015 [9], also see [10]. Interactive presentation of psychometric properties of the admission test is available in R package ShinyItemAnalysis [11].

The test consisted of 100 dichotomously scored items covering different Biology topics. For the purpose of this analysis, items were classified into three general domains – genetics, taxonomy, and human biology, respectively. The mutual proportions of these three domains were of nearly equal size.

To evaluate psychometric properties of the items, unidimensional two-parameter logistic (2PL) item-response theory (IRT) model was fitted to describe the probability of a correct answer given applicant's ability [12],

$$p_i(\theta_p) = \Pr(U_{pi} = 1|\theta_p, \boldsymbol{\xi}_i) = \Psi[a_i(\theta_p - b_i)] = \frac{\exp[a_i(\theta_p - b_i)]}{1 + \exp[a_i(\theta_p - b_i)]}, \quad (1)$$

where θ_p is the ability of subject $p \in \{1, 2, \ldots, N\}$, vector $\boldsymbol{\xi}_i = (a_i, b_i)^T$ stands for set of item parameters (discrimination and difficulty, respectively) for item $i \in \{1, 2, \ldots, I\}$, and $\Psi(\bullet)$ is the logistic function.

In the item calibration phase, the item parameters $(a_i, b_i)^T$ were estimated. To estimate the item parameters, we used marginal maximum likelihood (MML) as follows [13]. Let us assume local independence, i. e. independence of item responses for the same subject given their ability θ_p (within subject). Then the probability $\Pr(\mathbf{u}_p|\theta_p, \boldsymbol{\xi})$ of response pattern $\mathbf{u}_p$ of subject p follows the form

$$\Pr(\mathbf{u}_p|\theta_p, \boldsymbol{\xi}) = \prod_{i=1}^{I} \Pr(U_{pi} = u_{pi}|\theta_p, \boldsymbol{\xi}_i). \quad (2)$$

Supposing there is no cooperation between subjects, we can also assume independence between subjects (in-between). Let's further denote $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_I)$ the matrix of parameters for all items i. Then the marginal likelihood function takes form

$$L(\boldsymbol{\xi}, \mu, \sigma; \mathbf{U}) = \prod_{p=1}^{N} \Pr(\mathbf{u}_p|\boldsymbol{\xi}, \mu, \sigma) \quad (3)$$

with

$$\Pr(\mathbf{u}_p|\boldsymbol{\xi}, \mu, \sigma) = \int \ldots \int \Pr(\mathbf{u}_p|\theta_p, \boldsymbol{\xi})\, g(\theta_p|\mu, \sigma)\, \mathrm{d}\theta_p,$$

where μ and σ are the expected value and the variance of respondent ability θ_p. With this approach, abilities θ_p are treated as stochastic variables with normal distribution, $\theta_p \sim \mathcal{N}(\mu, \sigma)$ and are integrated out [14].

The first-order derivatives with respect to ability parameters θ_p result into the likelihood equations [15] that could be numerically estimated using Expectation-Maximization (EM) algorithm [16], producing the desired estimates of item parameters $\hat{\boldsymbol{\xi}} = \left(\hat{\boldsymbol{\xi}}_1, \ldots, \hat{\boldsymbol{\xi}}_I\right)$.

B. Settings of adaptive tests

Initialization. Initial item was selected as the one maximizing observed Fisher information at ability $\theta_0 = 0$, see [17].

Ability estimation. Test-taker's ability is iteratively updated whenever the respondent answers to a given item and the answer is collected. Beginning with the equation (2), the likelihood is as follows

$$L(\theta; \mathbf{u}_p) = \prod_{i=1}^{I} P(U_{pi} = u_{pi}|\theta, \boldsymbol{\xi}_i) \quad (4)$$

and is maximized with respect to θ. Then, the first-order and second-order partial derivatives are needed to compute

the maximum likelihood estimates (MLE) and their standard errors [18].

While we used MLE to estimate ability in most results shown here, other methods are available. In case the ability is only unidimensional, a popular approach is the weighted likelihood estimator (WLE) [19]; which maximizes equation (4) weighted by a function $w(\theta)$, thus

$$L\left(\theta; \mathbf{u}_p\right) = w(\theta) L\left(\theta|\mathbf{u}_p\right). \tag{5}$$

Finally, Bayesian ability estimation [20] specifies prior ability distribution $p\left(\theta_p|\mu,\sigma\right)$ and maximizes posterior distribution of θ_p given $\mathbf{u}_p$ of the following form:

$$p\left(\theta_p|\mathbf{u}_p,\boldsymbol{\xi},\mu,\sigma\right) = \frac{\Pr\left(\mathbf{u}_p|\theta_p,\boldsymbol{\xi}\right) p\left(\theta_p|\mu,\sigma\right)}{\int \ldots \int \Pr\left(\mathbf{u}_p|\theta_p,\boldsymbol{\xi}\right) p\left(\theta_p|\mu,\sigma\right) \mathrm{d}\theta_p}.$$

Item selection. We used likelihood-based item selection [17], i. e. in each step, the next (k-th) item was selected to maximize the observed Fisher information

$$i_k \equiv \arg\max_j \left\{ I_{\mathbf{U}_{k-1},U_j}\left(\hat{\theta}_{k-1}\right)\right\} \tag{6}$$

at $\theta_p = \hat{\theta}_{p,k-1}$ given a subject p, where $\mathbf{U}_{k-1}$ is an answer pattern up to the $(k-1)$-th item [17]. This rule is also known as the maximum-information rule in adaptive testing.

Other item selection procedures include naive approach such as Urry's criterion picking always an item with difficulty closest to the current ability estimate [21]. In Bayesian framework, the posterior distribution of θ_p after the preceding item serves as the prior distribution for the selection of the next item. If the posterior distribution after $k-1$ items has density $p\left(\theta|\mathbf{u}_{k-1}\right)$, then the k-th item is selected such that the posterior distribution

$$p\left(\theta|\mathbf{u}_{k-1}, U_{i_k}\right) \propto p\left(\theta|\mathbf{u}_{k-1}\right) p\left(U_{i_k} = u_{i_k}|\theta\right)$$

is optimized in some sense [20].

Termination criteria. In our simulation studies, we used ability estimate precision as a stopping rule. Assuming the $I_{\mathbf{U}_{p,k-1}}(\hat{\theta}_{p,k-1})$ is observed Fisher information [17] at $\hat{\theta}_{p,k-1}$ where $\mathbf{U}_{k-1}$ is an answer pattern until the $(k-1)$-th item (inclusively) given a subject p, then standard error of ability $\hat{\theta}_{p,k-1}$ is

$$\mathrm{SE}(\hat{\theta}_{p,k-1}) = \frac{1}{\sqrt{I_{\mathbf{U}_{p,k-1}}\left(\hat{\theta}_{p,k-1}\right)}}. \tag{7}$$

For the adaptive test, we specified the maximal allowed standard error $\mathrm{SE}(\theta)_{\max}$ of the ability estimate based on the distribution of standard errors of the ability estimates from the full 100-item test. For subject p, the test was terminated just after administration of the k-th item if $\mathrm{SE}(\hat{\theta}_{p,k}) \leq \mathrm{SE}(\theta)_{\max}$ and $\mathrm{SE}(\hat{\theta}_{p,k-1}) > \mathrm{SE}(\theta)_{\max}$. Otherwise, the test was stopped if the length of 100 items was reached and all available items were used.

Whenever the termination (stopping) criterion is met, the adaptive test is ended and final estimate of test-taker's ability is provided.

Content balancing. Balancing of an adaptive test content is usually treated as a combinatorial constrained optimization problem [22]. Alternatively, it is based on a shadow-test approach by projection of rest of the test at the current moment (after $k-1$ items are administrated), which is a nonlinear program using maximum-information rule and constrained by domain attributes and other conditions [22].

In the post-hoc analysis described in this paper, we used one of the combinatorial designs, where we initially set desired proportions of expected administration rate to each of the three domains (genetics, taxonomy, human biology). The items were selected in a way to minimize differences between the currently observed and initially set proportions.

Item exposure rate control. The rates of how many times each item is administrated to one or more of test-takers throughout one adaptive test session may be controlled to minimize their unwanted leakage outside the tested population. Hetter-Sympson experiment is commonly applied to face this problem and was also used in our simulation study [23]. The algorithm was run before the optimally selected item was administrated, output of which was a decision either to administer the item, or to pass and select the next best item at the current estimate of ability $\hat{\theta}_{p,k}$. The administered items were removed from the item pool. Hetter-Symspon experiment is based on evaluation of joint conditional probabilities of item administration; thus cumbersome and usually must be numerically simulated.

There are also some alternatives – an experiment determining which items are eligible for subjects and which not [24]. If an item is eligible, it remains in the pool for the subject p; otherwise it is removed. This works as a principle of "self-adjustment"; when an item was highly exposed within previous $p-1$ subjects, it is likely not to be eligible for the p-th subject.

C. Post-hoc analysis

In post-hoc analysis, the item parameters and the response patterns of the respondents were used to rerun the test under adaptive conditions. By doing this, the properties of the adaptive test (such as the test length, precision of estimated abilities etc.) were "post-hoc" evaluated and compared to the original linear test.

Considering the dataset of test-takers taking the real test, we varied the maximal allowed standard error of the ability estimates and ran the adaptive version of the test for each of the test-takers investigating how many items were needed to complete the test. The pseudocode of this simulation is provided in Algorithm 1.

Similarly, we calculated the z-score for each subject using the test scores from the real test,

$$z\text{-score} = \frac{x_p - \bar{x}}{s_x},$$

where x_p is a test score of a subject p, $\bar{x}$ is an average test score and s_x is a standard deviation of all test scores. All test-takers having their z-scores in the interval of $|z - z^*| \leq \delta$, where

Algorithm 1: Investigation of adaptive test length depending on the precision of ability estimate

Data: data of the real test
Result: boxplots of test lengths for CAT with different standard errors of ability estimates

```
{S}   // set or subset of respondents;
      // of the real test;
{A} = ∅  // list of vectors of lengths;
         // for different standard;
         // errors;
{E} = ∅  // list of standard errors;
for j = 1 : 7 do
    SE = 0.15 + 0.05 · j;
    {E} = {E} ∪ {SE};
    {D} = ∅ ;
    for p ∈ S do
        run an adaptive test for subject p with stopping
         criterion SE(θ)_max = SE and save its length
         as d;
        {D} = {D} ∪ {d};
    end
    {A} = {A} ∪ {D};
end
make a boxplot of {A} vs. {E} ;
```

Algorithm 2: Investigation of average adaptive test length depending on the z-score from the original linear test

Data: data of the real test
Result: boxplots of average test lengths for groups based on z-scores from the original linear test

```
δ = 0.05   // neighbourhood around z*;
{A} = ∅  // list of vectors of lengths;
         // for different z-scores;
{Z}      // list of original z-scores;
{Z*} = ∅ // list of z*-scores;
for j = 1 : 17 do
    z* = −2.00 + 0.25 · j;
    {Z*} = {Z*} ∪ {z*};
    {D} = ∅ ;
    for all subjects with z ∈ Z such that |z − z*| ≤ δ
     do
        run an adaptive test for the subject with
         stopping criterion SE(θ)_max = 0.30 and save
         its length as d;
        {D} = {D} ∪ {d};
    end
    {A} = {A} ∪ {D};
end
make a boxplot of {A} vs. {Z*} ;
```

$\delta = 0.05$ and $z^* \in \{-2.00, -1.75, -1.50, \ldots, +1.75, +2.00\}$ were supposed to virtually take the adaptive test, keeping the $\mathrm{SE}(\theta)_{\max} = 0.30$ for equation (7) constant. The z^* neighbourhood $\delta = 0.05$ was chosen empirically, but consequently, one can realize that $\delta = 0.125$ would cover continuously the entire range of all z-scores. For each z^*, a vector of all the adaptive tests' lengths was displayed in the final boxplot. The schema of the simulation is provided in Algorithm 2.

Similarly, the effect of content balancing and item exposure rate control was analyzed. When an adaptive test was administered to each test-taker from a randomly selected subset, we counted how many times individual items occur in the tests. Absolute numbers of the items' occurrences were then counted up for different scenarios – besides the situation when neither the content balancing nor the item exposure was applied, the case of (only) the content balancing and (only) the item exposure rate controlling was taken into account. Eventually, using the fact, the items were classified into three domains (Genetics, Taxonomy, Human Biology), their counts could be clearly plotted using boxes in a boxplot.

Finally, to study the impact of adaptive test with different settings on the admission process, we enumerated the admission mismatch rate between linear and adaptive tests. We assumed the best fifth of all the applicants would be admitted and we calculated the mismatch rate as the ratio of students who would be admitted based on their score in the linear test but not based on the score in adaptive test and vice versa. We then compared the admission mismatch rate for adaptive tests with stopping criteria $\mathrm{SE}(\theta)_{\max} \in \{0.20, 0.30, 0.40,\ 0.50\}$. While the best fifth for the linear test was calculated using the z-scores, the MML ability estimates were used for the adaptive test.

D. Monte Carlo simulation studies

Whereas the post-hoc analysis requires real data from an administrated test, the Monte-Carlo simulation study starts from the scratch – it generates abilities of "virtual" test-takers usually following normal distribution and responses based on selected model (e. g. the 2PL IRT model) with given item parameters. We first simulated the linear test, then, based on the simulated answers, the adaptive scenario was simulated. We then correlated ability estimates from the adaptive test with the true ability values. Finally, we displayed lengths of adaptive test and we correlated the ability estimate with the true ability. Other comparisons and analyses are possible (length with respect to the true ability score, etc.), but not presented here. The algorithm of the simulation is technically described in Algorithm 3.

Analyses were performed in R programming language and environment [25] using the package `mirtCAT` [26].

III. Results

All items of the linear test were calibrated using the 2PL IRT model as described by equation (1). Item characteristic curves and item information curves are plotted in Fig. 2 and Fig. 3. All items have positive discrimination $a_i > 0$ for

Algorithm 3: Investigation of ability estimates based on adaptive tests using a Monte-Carlo simulation

Data: generated abilities following $\mathcal{N}(0, 1^2)$, item parameters (estimated from real data), adaptive test's stopping criterion $\mathrm{SE}(\theta)_{\max} = 0.30$, item selection using maximum-information rule
Result: a list of ability estimates based on adaptive test

```
n = 300   // number of generated;
          // abilities;
{S}       // list of n generated;
          // abilities;
          // following N(0,1^2);
{A} = ∅   // list of adaptive-based;
          // ability estimates;
{D} = ∅   // list of lengths;
          // of adaptive tests;
for p = 1 : n do
    apply 2PL IRT model on p-th ability of {S} and
     simulate an answer pattern ;
    use the answer pattern and run an adaptive test for
     p-th ability and save its length as d and ability
     estimate as θ̂_p ;
    {A} = {A} ∪ θ̂_p;
    {D} = {D} ∪ {d};
end
make a boxplot of {D} ;
make a scatterplot of {A} vs. {S}, calculate
 a correlation of {A} and {S} ;
```

$\forall i \in \{1, 2, \ldots, 100\}$, resulting in a spectrum of the item information curves.

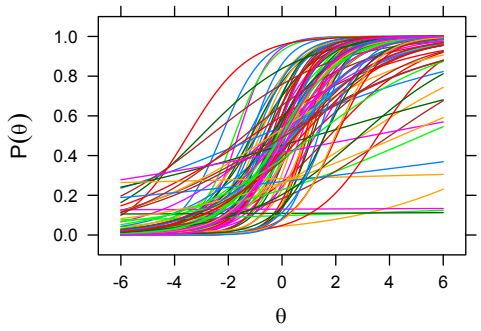

Fig. 2. Item characteristic curves of the linear test estimated using 2PL IRT model.

When applying the 2PL IRT model on the data from the linear test, we get, besides other, also standard errors of the ability estimates for each test-taker. Histogram of these standard errors in in Fig. 4. Range of the standard errors of the ability estimates is between 0.20 to 0.50, with majority of values within the interval $\langle 0.20, 0.30 \rangle$.

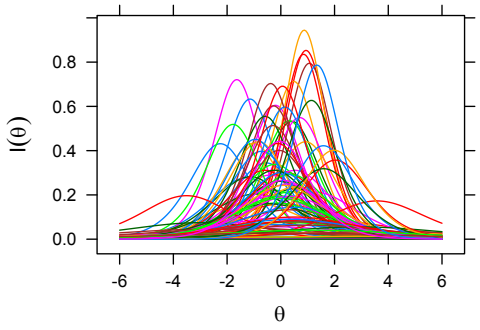

Fig. 3. Item information curves of the linear test estimated using 2PL IRT model.

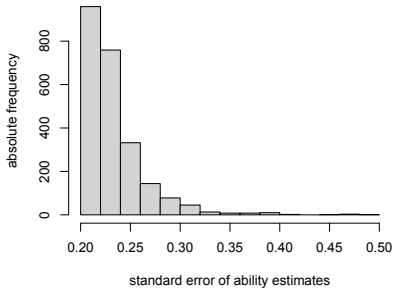

Fig. 4. Histogram of standard errors of the ability estimates.

A. Post-hoc analysis

Post-hoc analysis used the real test-takers data to simulate the results under scenario of an adaptive test with selected parameters. As an example, Fig. 5 demonstrates iteratively estimated ability estimates and order of items in which they would be administered to the 1-st subject under adaptive scenario with terminating criterion $\mathrm{SE}(\theta)_{\max} = 0.30$. We can see that the initial item would be item number 81, the last item would be item number 70. The width of the grey belt stands for precision of the ability estimate at each step k, equal to two standard errors $2\mathrm{SE}(\hat{\theta})_{p,k}$ of the ability estimate of person p. The belt becomes more narrow as the test-taker answers more and more items. Note that the standard error after 18 administered items is $\mathrm{SE}(\hat{\theta})_{1,18} \leq 0.30$ while after 17 administered items it is $\mathrm{SE}(\hat{\theta})_{1,17} > 0.30$.

As a result of Algorithm 1, Fig. 6 presents how the number of items needed to stop the adaptive test depends on the termination criterion. We can see that the higher the maximal standard error is applied as the termination criterion, the lower the number of items is needed to terminate the adaptive test.

As a result of simulation described with Algorithm 2, Fig. 7 illustrates how the respondent ability (estimated with a z-score) affects the number of items needed to stop the adaptive test. The size of maximal allowed standard error of the ability estimates as the stopping criterion was set to $\mathrm{SE}(\theta)_{\max} = 0.30$ based on the distribution of the standard

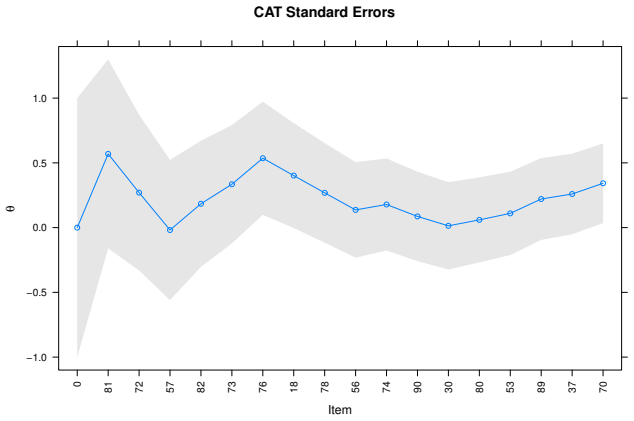

Fig. 5. A plot of progress of 1-st subject in an adaptive test with the terminating criterion set to maximal allowable standard error of the ability estimates of $\mathrm{SE}(\theta)_{\max} = 0.30$.

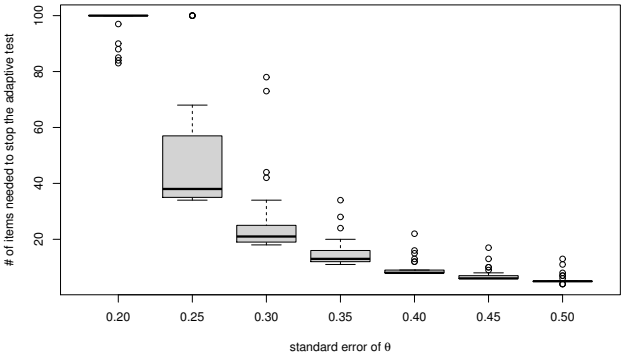

Fig. 6. Number of items needed to stop the adaptive test versus a size of standard error of the ability estimate as the stopping criterion.

errors, displayed in Fig. 4. We can see that the closer the z-score is to zero, the lower number of items is needed to complete the adaptive test while meeting the required ability estimate precision defined by $\mathrm{SE}(\theta)_{\max} = 0.30$. This corresponds to the fact that the information functions for majority of items have the maxima for ability around zero as demonstrated in Fig. 3. Contrary, for z-scores far from zero, the observed Fisher information is small for most of the items, thus a larger number of items is needed to meet the stopping criterion, and often not even meeting it using all 100 items available.

In Fig. 8, we plot numbers of occurrences of items in all individual adaptive tests for randomly selected 50 test-takers, considering that each item belongs to one of the following three domains – either to genetics, taxonomy, or human biology, respectively. While the proportions of the three domains of items as they were administrated vary a lot in Fig. 8 where neither the content balancing nor the item exposure rate control is applied, these numbers are near equal when the content balancing is applied. When the item exposure rate control

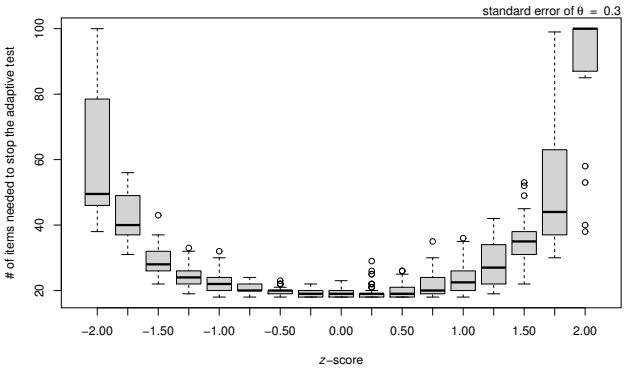

Fig. 7. Number of items needed to stop the adaptive test in respondents of different ability levels.

is employed, there is no visible change in comparison to no application of the exposure control.

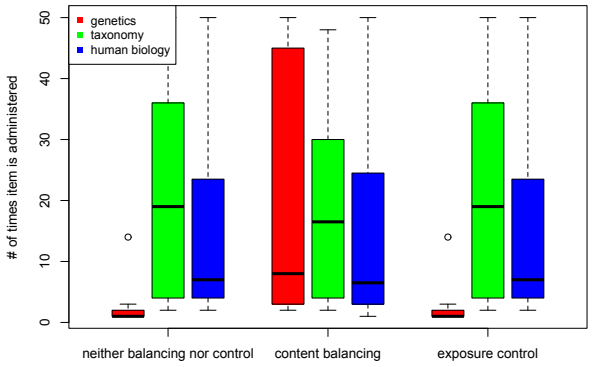

Fig. 8. Number of items belonging to the domains *genetics*, *taxonomy*, *human biology*, respectively, as were administrated with application of neither content balancing nor item exposure control, with application of content balancing, and with application of item exposure rate control only.

Table I provides mismatch matrices for linear and adaptive tests with different stopping criteria $\mathrm{SE}(\theta)_{\max}$. As expected, the mismatch rate increases with increased allowed standard error applied as a stopping criterion in the adaptive test. The mismatch rate is 0.036, 0.083, 0.102 and 0.118 for adaptive tests with stopping rules $\mathrm{SE}(\theta)_{\max} = 0.20$, 0.30, 0.40 and 0.50, respectively.

B. Monte-Carlo simulation study

As a result of the Monte-Carlo simulation study described by Algorithm 3, Fig. 9 provides a boxplot illustrating the mean length of the adaptive test for the set of test-takers with the generated abilities. While each test-taker has to answer to all (100) items within the linear fashion, they would only have to answer about 25 % of items to finish the simulated adaptive test with the termination criterion $\mathrm{SE}(\theta)_{\max} = 0.30$. The length of the test using this adaptive scenario provides

TABLE I
MISMATCH MATRICES OF ADMITTED TEST-TAKERS BY LINEAR AND ADAPTIVE TEST WITH STOPPING CRITERION
$\mathrm{SE}(\theta)_{\max} \in \{0.20, 0.30, 0.40, 0.50\}$.

$\mathrm{SE}(\theta)_{\max} = 0.20$		admitted by adaptive test	
		no	yes
admitted by linear test	no	1842	38
	yes	48	435
$\mathrm{SE}(\theta)_{\max} = 0.30$		admitted by adaptive test	
		no	yes
admitted by linear test	no	1787	93
	yes	103	380
$\mathrm{SE}(\theta)_{\max} = 0.40$		admitted by adaptive test	
		no	yes
admitted by linear test	no	1762	118
	yes	123	360
$\mathrm{SE}(\theta)_{\max} = 0.50$		admitted by adaptive test	
		no	yes
admitted by linear test	no	1737	143
	yes	136	347

75% shortening as compared to the linear test, while keeping the same precision of ability estimates for most respondents.

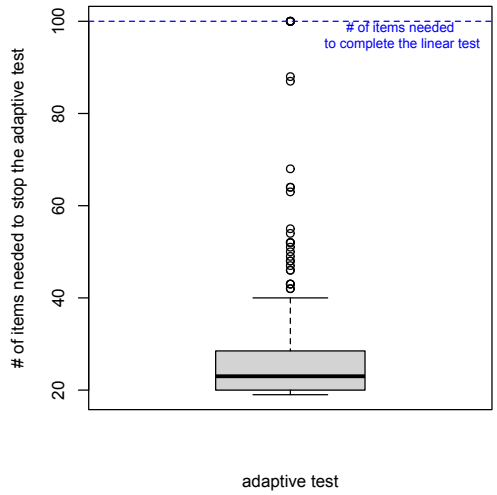

Fig. 9. A boxplot of number of items needed to be answered to complete the adaptive test based on Monte-Carlo simulated test-takers' abilities. The blue dashed line shows a length of the linear test (100 items).

Pearson's correlation between the generated abilities and their estimates based on the adaptive tests is about $\rho \doteq 0.960$, which is depicted also in Fig. 10.

IV. CONCLUSION

Both the post-hoc analysis and Monte-Carlo simulation study showed that average test lengths can be shortened with adaptive tests, while keeping the standard error of the ability estimates at the same level for most of the respondents. The shortening of the test within the adaptive test with $\mathrm{SE}(\hat{\theta})_{\max} = 0.30$ was by about 75 % percent, i. e. while the original linear test had 100 items, the adaptive one was ended on average after answering 25 items only.

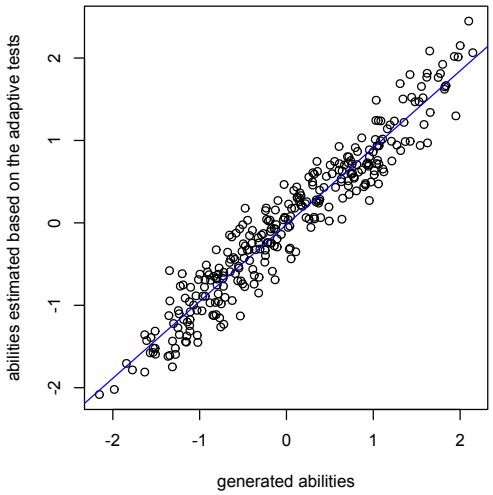

Fig. 10. A scatterplot of the generated abilities and their estimates based on the adaptive tests. The blue line stands for an axis of the first quadrant of the plot.

When even larger standard errors of the ability estimates are tolerated, the length of the test could be reduced even more, e. g. to only 10 items per one test, as was shown in the post-hoc analysis of the average adaptive test length with varying stopping criterion.

The post-hoc simulation also demonstrated that an average length for adaptive tests is shorter for average ability levels.

While the content balancing with the combinatorial approach showed a significant improvement in test domain equalizing, an effect of the item exposure rate did not seem to be so eminent under our setting.

The lower the tolerated standard error as a stopping criterion of the adaptive test is, the lower is the mismatch error rate when using an adaptive test instead of the linear one. The mismatch rate was less than 10% for adaptive test with stopping criterion of $\mathrm{SE}(\hat{\theta})_{\max} = 0.30$.

The Monte-Carlo simulation study also indicated that ability estimates provided by the adaptive tests can be tightly correlated with their true (generated) values; thus, although the shortened length, the adaptive test can provide precise estimates of the respondent abilities.

To conclude, usage of adaptive testing seems to be a promising alternative to classic linear tests and offers many advantages as showed by the simulations.

ACKNOWLEDGMENT

Research was supported by Charles University grant PRIMUS/17/HUM/11.

REFERENCES

[1] Wim J Linden, Wim J van der Linden, and Cees AW Glas. *Computerized adaptive testing: Theory and practice*. Springer, 2000.

[2] Howard Wainer, Neil J Dorans, Ronald Flaugher, et al. *Computerized adaptive testing: A primer*. Routledge, 2000.

[3] David Magis, Duanli Yan, and Alina A Von Davier. *Computerized adaptive and multistage testing with R: Using packages catr and mstr*. Springer, 2017.

[4] David J Weiss and G Gage Kingsbury. "Application of computerized adaptive testing to educational problems". In: *Journal of Educational Measurement* 21.4 (1984), pp. 361–375.

[5] Jan Stochl, Jan R Böhnke, Kate E Pickett, et al. "Computerized adaptive testing of population psychological distress: simulation-based evaluation of GHQ-30". In: *Social psychiatry and psychiatric epidemiology* 51.6 (2016), pp. 895–906.

[6] Jan Stochl, Jan R Böhnke, Kate E Pickett, et al. "An evaluation of computerized adaptive testing for general psychological distress: combining GHQ-12 and Affectometer-2 in an item bank for public mental health research". In: *BMC medical research methodology* 16.1 (2016), p. 58.

[7] Dagmar Amtmann, Alyssa M Bamer, Jiseon Kim, et al. "A comparison of computerized adaptive testing and fixed-length short forms for the Prosthetic Limb Users Survey of Mobility (PLUS-MTM)". In: *Prosthetics and orthotics international* 42.5 (2018), pp. 476–482.

[8] Karon F Cook, Seung W Choi, Paul K Crane, et al. "Letting the CAT out of the bag: comparing computer adaptive tests and an eleven-item short form of the Roland-Morris Disability Questionnaire". In: *Spine* 33.12 (2008), p. 1378.

[9] Patricia Martinková, Lubomír Štěpánek, Adéla Drabinová, et al. "Semi-real-time analyses of item characteristics for medical school admission tests". In: *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems*. Ed. by M. Ganzha, L. Maciaszek, and M. Paprzycki. Vol. 11. Annals of Computer Science and Information Systems. IEEE, 2017, pp. 189–194. DOI: 10.15439/2017F380. URL: http://dx.doi.org/10.15439/2017F380.

[10] Čestmír Štuka, Patrícia Martinková, Karel Zvára, et al. "The prediction and probability for successful completion in medical study based on tests and pre-admission grades". In: *New Educational Review* 28 (2012), pp. 138–52.

[11] Patrícia Martinková and Adéla Drabinová. "ShinyItemAnalysis for Teaching Psychometrics and to Enforce Routine Analysis of Educational Tests." In: *R Journal* 10.2 (2018).

[12] Wim J. van der Linden and Cees A.W. Glas. "25 Statistical Aspects of Adaptive Testing". In: *Handbook of Statistics*. Elsevier, 2006, pp. 801–838. DOI: 10.1016/s0169-7161(06)26025-5. URL: https://doi.org/10.1016/s0169-7161(06)26025-5.

[13] R. Darrell Bock and Murray Aitkin. "Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm". In: *Psychometrika* 46.4 (Dec. 1981), pp. 443–459. DOI: 10.1007/bf02293801. URL: https://doi.org/10.1007/bf02293801.

[14] Yoshio Takane and Jan de Leeuw. "On the relationship between item response theory and factor analysis of discretized variables". In: *Psychometrika* 52.3 (Sept. 1987), pp. 393–408. DOI: 10.1007/bf02294363. URL: https://doi.org/10.1007/bf02294363.

[15] Cees A. W. Glas. "Modification indices for the 2-PL and the nominal response model". In: *Psychometrika* 64.3 (Sept. 1999), pp. 273–294. DOI: 10.1007/bf02294296. URL: https://doi.org/10.1007/bf02294296.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society, Series B* 39.1 (1977), pp. 1–38.

[17] Hua-Hua Chang and Zhiliang Ying. "Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests". In: *The Annals of Statistics* 37.3 (June 2009), pp. 1466–1488. DOI: 10.1214/08-aos614. URL: https://doi.org/10.1214/08-aos614.

[18] Daniel O. Segall. "Multidimensional adaptive testing". In: *Psychometrika* 61.2 (June 1996), pp. 331–354. DOI: 10.1007/bf02294343. URL: https://doi.org/10.1007/bf02294343.

[19] Thomas A. Warm. "Weighted likelihood estimation of ability in item response theory". In: *Psychometrika* 54.3 (Sept. 1989), pp. 427–450. DOI: 10.1007/bf02294627. URL: https://doi.org/10.1007/bf02294627.

[20] Frederic Lord. *Applications of item response theory to practical testing problems*. Hillsdale, N.J: L. Erlbaum Associates, 1980. ISBN: 978-0898590067.

[21] Frank L. Schmidt, John E. Hunter, and Vern W. Urry. "Statistical power in criterion-related validation studies." In: *Journal of Applied Psychology* 61.4 (1976), pp. 473–485. DOI: 10.1037/0021-9010.61.4.473. URL: https://doi.org/10.1037/0021-9010.61.4.473.

[22] Wim J. van der Linden and Richard M. Luecht. "Observed-score equating as a test assembly problem". In: *Psychometrika* 63.4 (Dec. 1998), pp. 401–418. DOI: 10.1007/bf02294862. URL: https://doi.org/10.1007/bf02294862.

[23] Rebecca D. Hetter and J. Bradford Sympson. "Item exposure control in CAT-ASVAB." In: *Computerized adaptive testing: From inquiry to operation.* American Psychological Association, 1997, pp. 141–144. DOI: 10.1037/10244-014. URL: https://doi.org/10.1037/10244-014.

[24] Martha L. Stocking and Charles Lewis. "Controlling Item Exposure Conditional on Ability in Computerized Adaptive Testing". In: *Journal of Educational and Behavioral Statistics* 23.1 (1998), p. 57. DOI: 10.2307/1165348. URL: https://doi.org/10.2307/1165348.

[25] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017. URL: https://www.R-project.org/.

[26] R. Philip Chalmers. "Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications". In: *Journal of Statistical Software* 71.5 (2016), pp. 1–39. DOI: 10.18637/jss.v071.i05.