LSAC RESEARCH REPORT SERIES

Outlier Detection in High-Stakes College Entrance Testing

Rob R. Meijer
University of Twente, Enschede, The Netherlands

Law School Admission Council
Computerized Testing Report 01-08
September 2005

A Publication of the Law School Admission Council

The Law School Admission Council (LSAC) is a nonprofit corporation whose members are more than 200 law schools in the United States and Canada. It was founded in 1947 to coordinate, facilitate, and enhance the law school admission process. The organization also provides programs and services related to legal education. All law schools approved by the American Bar Association (ABA) are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also included in the voting membership of the Council.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the author and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

Though the development of computerized adaptive testing (CAT) has resulted in more efficient educational and psychological measurement, it has also generated new practical and theoretical problems. One theoretical problem that arises is the identification of item score patterns (correct and incorrect responses) for particular test takers that do not conform to what would be expected based on the mathematical model being applied. An example of such an aberrant item score pattern would be that of a test taker who answered many easy items incorrectly and many difficult items correctly. If a test taker has an item score pattern that does not fit, the pattern is unlikely to give valuable information about the test taker's ability, but may point toward other behavior during the test. Explanations of aberrant item score patterns in CAT differ from those in paper-and-pencil testing. For example, in paper-and-pencil testing, answer copying may result in unexpected item scores, whereas in a CAT, direct answer copying is impossible because different test takers are administered different tests.

Several recent person-fit methods for CAT were studied. The sampling distributions of these statistics were studied empirically, and it was shown that the statistics have sufficient power to relate different patterns of responses to different types of misfit in the test taker's behavior.

Abstract

In this study we discuss recent developments of person-fit analysis in the context of computerized adaptive testing (CAT). Methods from statistical process control are discussed that have been proposed to classify an item score pattern as fitting or misfitting the underlying item response theory (IRT) model in a CAT. Most person-fit research in CAT is restricted to simulated data. In this study, empirical data from a high-stakes test are used. Alternative methods to generate norm distributions to allow the determination of bounds are discussed. These bounds may be used to classify item score patterns as fitting or misfitting. Using bounds determined from the sample, the empirical analysis indicated that different types of misfit can be distinguished. Possibilities to use this method as a diagnostic instrument are discussed.

Introduction

In computerized adaptive testing (CAT), for each examinee the ability is estimated during test administration continually, and subsequent items are selected to match the current ability estimate. CAT originated from the idea that matching item difficulty and ability level will result in a more efficient and reliable way of testing. Item response theory (IRT) models (e.g., Embretson & Reise, 2000) model response behavior with distinct parameters for the person's ability and the item characteristics. IRT models allow the construction of different tests for different examinees from an item bank and compare their results on the same latent trait scale. For an introduction to CAT and an overview of recent developments, refer to Wainer (1990), van der Linden and Glas (2000), and Meijer and Nering (1999).

The development of CAT has resulted in a more efficient way of educational and psychological testing and new innovations that make testing more reliable and valid. CAT has also generated new practical and theoretical problems. In this study we will focus on the fit of an item score pattern to an IRT model in CAT, a research topic that has been underexposed in the literature. For paper-and-pencil (P&P) tests, a large number of studies have focused on this topic; for a review refer to Meijer and Sijtsma (1995; 2001). The central idea in these studies is that, although the items in a test may show a reasonable fit to an IRT model, an individual's score pattern may be very unlikely given the IRT model. In educational and psychological testing the main aim is to measure persons, and any indication that a person's score pattern is very unlikely given the model is valuable information. It may point at other mechanisms than the assumed interaction between the trait and item characteristics that is modeled via an IRT model. Therefore, both for P&P tests and CAT, identifying misfitting item score patterns is important, although the cause of misfit may be different for both types of tests. For example, in P&P tests, answer-copying may result in unexpected item scores, whereas in a CAT, answer copying is improbable because different examinees get different tests.

Let us give two examples to illustrate the importance of investigating the fit of an item score pattern in a CAT and to illustrate the mechanisms underlying aberrant response behavior. Kingsbury and Houser (1999) describe the situation where a CAT is routinely used as a pretest and posttest to check if short-term changes in instruction in a curriculum has any effect on the mean achievement level of the candidates. In this situation, some students may not take the test seriously and may guess the answers to some or all of the items. This disinterest in the results of the test may result in item score patterns that are unexpected based on the IRT model that is being used. Note that the number-correct score for these persons on the test may also be lower than the score they would have obtained if they answered the items according to their own proficiency. When many persons in this situation do not take the test seriously, the incorrect conclusion

would be drawn that the curriculum adaptation will result in lower number-correct scores. In general, this problem exists in all situations in which a test is used as an instrument to assess the quality of a curriculum and where the results are not primarily used to evaluate the students.

As another example, consider high-stakes testing, where it is important that a test agency guarantee that the test has the same psychometric characteristics for each person. Because in CAT different (sub)tests are given to different persons, it is assumed that the ability is invariant over subtests of the total test, so that different subtests can be administered to different persons. Violations of invariant ability therefore may be a serious threat to the comparability of the test scores across persons. Routinely checking the invariant ability level for each person is therefore important.

Because of the idiosyncrasies of CAT compared to P&P tests, we will (1) discuss the possibility of using existing person-fit statistics in a CAT, (2) discuss a number of recently proposed person-fit methods for CAT and discuss their pros and cons, and (3) conduct an empirical study in which we apply one of these methods to an empirical dataset.

Person-Fit Research

Several statistics have been proposed to investigate the fit of an item score pattern to an IRT model. In IRT the probability of obtaining a correct answer on item I ($I = 1, ..., k$) is a function of the latent trait value (θ) and the characteristics of the item such as the location b (van der Linden & Hambleton, 1997; Embretson & Reise, 2000). This conditional probability $p_i(\theta)$ is the item response function (IRF). Let x_i denote the binary (0, 1) score to item I and let $\mathbf{x} = (x_1, ..., x_k)$ denote the binary response vector, a_i the item discrimination parameter, b_i the item difficulty parameter, and c_i the item guessing parameter. The probability of correctly answering an item according to the three-parameter logistic IRT model (3PLM) is defined by

$$P_g(\theta) = c_i + \frac{(1 - c_i)\exp\left[a_i(\theta - b_i)\right]}{1 + \exp\left[a_i(\theta - b_i)\right]}, \tag{1}$$

when $c_i = 0$ for all items the 3PLM becomes the two-parameter logistic IRT model (2PLM).

Most person-fit research has been conducted using fit statistics that are designed to investigate the probability of an item score pattern under the null hypothesis of fitting response behavior. A general form in which most person-fit statistics can be expressed is

$$W = \sum\left[x_i - p_i(\theta)\right]w_i(\theta) \tag{2}$$

where the expectation of the statistic equals 0. For example, Wright and Stone (1979) proposed a person-fit statistic

$$U = \sum_{i=1}^{k} \frac{\left[x_i - p_i(\theta)\right]^2}{kp_i(\theta)\left[1 - p_i(\theta)\right]}. \tag{3}$$

Most studies (e.g., Levine and Rubin, 1979; Drasgow, Levine, & Williams, 1985) have been conducted using some suitable function of the log-likelihood function

$$l = \sum_{i=1}^{k}\left\{x_i \ln p_i(\theta) + (1 - x_i)\ln\left[1 - p_i(\theta)\right]\right\}. \tag{4}$$

Large negative values of this statistic indicate misfitting response behavior. Often a standardization of l is used with an expectation of 0 and a variance of 1.

For relatively short P&P tests, the variance of this statistic is underestimated (i.e., less than 1), and corrections can be applied (e.g., Snijders, 2001). Using statistics like U and l in a CAT is problematic. This will be illustrated on the basis of the item score patterns depicted in Table 1. In Table 1 all possible item score patterns on a test are depicted with their value on the person-fit statistic M proposed by Molenaar and Hoijtink (1990). This statistic is equivalent to l using the Rasch (1960) model. Two sets of M-values are depicted; the first set is based on the item difficulty values (–2, –1, 0, 1, 2) and the second set is based on the

item difficulty values (–1, –0.5, 0, 0.5, 1). For both item difficulty sets, pattern #1 is the most plausible pattern and pattern #7 is the least plausible pattern. It can be seen that reducing the variance in the item difficulties also results in a reduced variance of *M*. In the extreme case when all items have the same item difficulty, all item score patterns have the same likelihood.

TABLE 1
M-values for different item score patterns

	Items					M-values	
Pattern	1	2	3	4	5	Difficulty Set 1	Difficulty Set 2
1	1	1	0	0	0	3	1.5
2	1	0	1	0	0	2	1.0
3	1	0	0	1	0	1	0.5
4	1	0	0	0	1	0	0.0
5	0	1	0	0	1	–1	–0.5
6	0	0	1	0	1	–2	–1.0
7	0	0	0	1	1	–3	–1.5

Difficulty Set 1 has difficulty values –2, –1, 0, 1, and 2.
Difficulty Set 2 has difficulty values –1, –0.5, 0, 0.5, and 1.
$M = -\sum b_i x_i$

The situation with reduced variance in the item difficulties is relevant for person-fit in a CAT. Due to the relatively modest variability in the item difficulties in a CAT compared to those in a P&P test, fitting and misfitting item score patterns are difficult to distinguish.

This was illustrated using simulated data by van Krimpen-Stoop and Meijer (1999; see also Reise, 1995) who showed that in CAT the distributional characteristics of existing person-fit statistics like *l* are not in agreement with their theoretical distributions. They found that empirical type I errors are too small compared to nominal type I errors.

Person Fit in CAT

In a few studies person-fit statistics have been proposed to be used in a CAT. McLeod & Lewis (1999) proposed a statistic Z_c that was designed to detect item score patterns that result from memorization. Before the statistic can be calculated, the item bank is divided into three parts; easy items, items of medium difficulty, and difficult items. Let $Easy[p_i(\theta) - x_i]$ denote the mean residual for the easy items and $Diff[p_i(\theta) - x_i]$ the mean residual for the most difficult items in an administered CAT, then Z_c is given by

$$Z_c = \frac{Easy\left[p_i(\theta) - x_i\right] - Diff\left[p_i(\theta) - x_i\right]}{\left\{\sum Easy\left\{p_i(\theta)\left[1 - p_i(\theta)\right]\right\} / n^2_{Easy}\right\} + \left\{\sum Diff\left\{p_i(\theta)\left[1 - p_i(\theta)\right]\right\} / n^2_{Diff}\right\}}; \quad (5)$$

Z_c is positive when an examinee answered the easy items incorrectly and the difficult items correctly. Applying this statistic to an operational Graduate Record Examination Quantitative CAT bank with 14% simulated memorized items resulted, however, in low detection rates. Z_c was constructed to detect examinees with preknowledge of the item scores. A drawback of Z_c is that each examinee should receive at least one easy and one difficult item and, thus, the item selection algorithm should be adapted when using this fit statistic. In addition, the moderately difficult items of a person are not taken into account, which results in an incomplete picture of the fit of an examinee's item score pattern. Another drawback, similar to person-fit statistics proposed in the context of P&P tests, is that the statistic does not identify where the misfitting items appear in the order of presentation. As will be seen later, the proposed statistic does provide this important information.

Drasgow, Levine, and Zickar (1996) discussed a method to detect random response behavior in a CAT. This method was based on a likelihood ratio test comparing the likelihood of an item score pattern under the IRT model with the likelihood of the item score pattern under an alternative model (Drasgow & Levine, 1986). As an alternative model, random response behavior was modeled by a two-stage process. In the first stage, it was assumed that as a result of unfamiliarity with the computer, examinees devoted all their intellectual resources to learning how to interact with the computer. Consequently, their responses to the first n_1 items can be viewed as essentially random. Then, it was assumed that examinees mastered the mechanics of responding to a computer test by the time n_1 items were administered. Thus, it was assumed that the final n_2 items would be answered according to the model for normal responding.

Given this model, the conditional likelihood of the response pattern $x = (x_1, x_2)$ for the misfitting model is

$$P_{misfit} = (x \mid \theta) = \prod_{i=1}^{k} \left(\frac{1}{g}\right)^{x_i} \left(\frac{g-1}{g}\right)^{1-x_i} P_{fit}(x_2 \mid \theta) \tag{6}$$

where $1/g$ is taken as the probability of a correct response during the subtest of n_1 items and $g - 1/g$ is the probability of an incorrect response. The marginal likelihood can then be obtained by integration with respect to the density. An important drawback of this method is that in practice it is difficult to come up with a plausible alternative model. How many items will the examinee guess the answers to? Will he/she completely guess such that the probability of a correct answer is 1/g?

As an alternative to both methods discussed above, both Bradlow, Weiss, & Cho (1998) and van Krimpen-Stoop and Meijer (2000) proposed person-fit statistics in which a model-fitting item-score pattern consists of an alternation of correct and incorrect responses, especially at the end of the test when $\hat{\theta}$ converges on θ. A string of consecutive correct or incorrect answers could signal misfit. Sums of consecutive negative or positive residuals $[x_i - p_i(\theta)]$ can be investigated using a cumulative sum procedure (Page, 1954). For each item I in the test, a statistic T_i can be calculated that equals a weighted version of $[x_i - p_i(\theta)]$. A simple statistic is

$$T_1 = 1/k[x_i - p_i(\theta)]. \tag{7}$$

Then, the sum of these T_is is accumulated as follows:

$$C_i^+ = \max\left[0, T_i + C_{i-1}^+\right], \tag{8}$$

$$C_i^- = \min\left[0, T_i + C_{i-1}^-\right], \text{and} \tag{9}$$

$$C_0^+ = C_0^- = 0, \tag{10}$$

where C^+ and C^- reflect the sum of consecutive positive and negative residuals, respectively. Let *UB* and *LB* be some appropriate upper and lower bounds. Then, when $C^+ > UB$ or $C^- < LB$, the item-score pattern can be classified as not fitting the model; otherwise, the item score pattern can be classified as fitting.

To illustrate the use of this statistic, consider a 20-item CAT with items selected from a simulated 400-item pool fitting the 2PLM with item parameters $a_i \sim N$ (1, 0.2) and $b_i \sim U$ (–3, 3). Simulation results (van Krimpen-Stoop & Meijer, 2000) showed that, when the CUSUM was determined using the statistic *T*, the values of *UB* and *LB* at $\alpha = .05$ were .13 and –.13, respectively. To investigate the fit of an item score pattern, first $\hat{\theta}$ is determined. In a CAT, two different values of $\hat{\theta}$ can be chosen to calculate *T*: the value of the updated $\hat{\theta}_{k-1}$, or the final $\hat{\theta}_N$. Using $\hat{\theta}_{k-1}$, the fit can be investigated during test administration; however, $\hat{\theta}_N$ is more accurate and results in more stable results (van Krimpen-Stoop & Meijer, 2000). Therefore, $\hat{\theta}_N$ is used in this example. *T* is determined for each administered item. Based on the values of *T* and according to Equations (7) through (9), C^+ and C^- are calculated for each administered item.

Consider an examinee taking the 20-item CAT generating the item score pattern given in Table 2. The final ability estimate, $\hat{\theta}_N$, for this examinee was –0.221. Table 2 gives the values T_1, C^+, and C^- after the administration of each item. Consider the first three items. The first item score equals 0 and $p_i(\theta) = .411$; this results in $T_1 = -.021$ (Equation 6). Substituting this value in (7) results in $C_1^+ = 0$ and in (8) results in $C_1^- = -.021$. Answering the second item incorrectly results in $T_2 = -.022$, $C_2^+ = 0$, and $C_2^- = -.042$. The third item is answered correctly and thus $T_3 = .025$, $C_3^+ = 0 + .025 = .025$ and $C_3^- = -.042 + .025 = -.017$.

TABLE 2
CUSUM procedure for a fitting item score pattern

Item	X	P	T	C^+	C^-
1	0	.411	–.021	0	–.021
2	0	.439	–.022	0	–.042
3	1	.497	.025	.025	–.017
4	0	.476	–.024	.001	–.041
5	1	.580	.021	.022	–.020
6	0	.463	–.023	0	–.043
7	1	.514	.024	.024	–.019
8	0	.578	–.029	0	–.048
9	1	.664	.017	.017	–.031
10	1	.568	.022	.038	–.009
11	1	.534	.023	.062	0
12	0	.287	–.014	.047	–.014
13	0	.424	–.021	.026	–.036
14	1	.557	.022	.048	–.013
15	0	.411	–.021	.028	–.034
16	0	.421	–.021	.007	–.055
17	1	.679	.016	.023	–.039
18	1	.418	.029	.052	–.010
19	0	.319	–.016	.036	–.026
20	1	.606	.020	.056	–.006

Note that the procedure is running on both sides and that a negative (or positive) value contributes both to C^+ or C^-. Because 0 is the smallest value C^+ can obtain and the largest value C^- can obtain, we can distinguish strings of positive and negative residuals. For this particular item score pattern, it can be seen (Table 2, columns 5 and 6) that C^+ stays below .13 and C^- stays above –.13. The highest value of C^+ is .062 (item 11), and the lowest value of C^- is –.055 (item 16). Therefore, this item score pattern is classified as *fitting* at $\alpha = .05$.

Consider now an examinee who responds to the 20-item CAT by randomly guessing the correct answers to all the items with 5 alternatives per item. In Table 3, the item score pattern for this examinee ($\hat{\theta}_N = -3.5$) and T, C^+, and C^- are given. C^+ stays below .13, whereas at the 19th item the value of C^- becomes smaller than –.13. As a result, the item score pattern is classified as *misfitting*.

TABLE 3
CUSUM procedure for a misfitting (guessing) item score pattern

Item	X	P	T	C^+	C^-
1	0	.005	0	0	0
2	0	.006	0	0	–.001
3	1	.007	.050	.050	0
4	0	.009	0	.049	0
5	0	.012	–.001	.049	–.001
6	0	.026	–.001	.047	–.002
7	0	.060	–.003	.044	–.005
8	0	.070	–.003	.041	–.009
9	1	.134	.043	.084	0
10	0	.082	–.004	.080	–.004
11	0	.149	–.007	.073	–.012
12	0	.251	–.013	.060	–.024
13	0	.277	–.014	.046	–.038
14	0	.259	–.013	.033	–.051
15	0	.330	–.017	.017	–.067
16	0	.304	–.015	.002	–.083
17	0	.359	–.018	0	–.101
18	0	.360	–.018	0	–.119
19	0	.364	–.018	0	–.137
20	1	.305	.035	.035	–.102

Researchers van Krimpen-Stoop and Meijer (2000) used simulated data to investigate the power of this statistic under different types of misfitting behavior. They simulated random response behavior (chance), violations of local independence, and noninvariant ability. In the latter case, two different θ values for each simulee were used to generate responses. At an α-level of .05, they reported detection rates of .60 under the guessing condition and between .22 and .72 under the noninvariant ability condition depending on the differences between the thetas. The detection rates under the guessing condition were low at .10.

As noted, van Krimpen-Stoop and Meijer (2000) used simulated data to determine the power of the statistics. We will apply the statistic to empirical data.

Method

Data

To compare and investigate the usefulness of different person-fit methods, we analyzed the empirical data of a high-stakes test. The minimum test length is 70 items, and the maximum test length is 140 items. If at the end of the administration of 70 items a pass/fail decision is reached with 95% confidence, the examination ends. If a pass/fail decision cannot be reached with 95% confidence, the examination continues until a pass/fail decision can be made with 95% confidence, or until the individual has taken 140 items, or until the time limit of 3 hours is reached. The content is balanced according to a blueprint, and data are calibrated according to the Rasch model. Each test contains five different subject matters. Furthermore, the first item is administered near the pass point ($\theta = 1$), and the first 10 items are administered within .10 logits of the previously administered item difficulty.

Sampling Distribution

To decide if an item score pattern is very unusual under the model, a distribution of the statistics is needed. In CAT there are different possibilities to choose the distribution $f(\mathbf{x})$ (Bradlow et al., 1998). The main question is on what information we want to condition. In this study we chose the simplest alternative; that is, we used the distribution $f(\mathbf{x} \mid \hat{\theta})$ where we assumed that the item difficulties were known and fixed. Then, when we tested at, for example, a 5% level, we first determined for each $\mathbf{x}$ the most extreme value and determined the *LB* and *UB* by choosing that value for which 2.5% of the most extreme values lie above (*UB*) or below (*LB*) that value across $\mathbf{x}$. Because the test we analyzed has a variable length, we did not condition on test length. Some closer inspection of the results for different test length revealed that there was no effect of test length on the *LB* and the *UB*.

An alternative would be to take the stochastic nature of $\hat{\theta}$ into account and sample from the posterior predictive density, that is to determine

$$f(\mathbf{x} \mid \mathbf{x}_{obs}) = \int f(\mathbf{x} \mid \theta) p(\theta \mid \mathbf{x}_{obs}) d\theta,$$

or to sample from the prior predictive distribution

$$f(\mathbf{x}) = \int f(\mathbf{x} \mid \theta) p(\theta) d\theta.$$

In general, sampling from the posterior or prior predictive distribution with a normal prior may have the drawback that the fit of an item score pattern is compared with the fit of a person with average θ. In the person-fit literature for P&P tests it is shown that the distribution of a person-fit statistic may depend on θ. Sampling from the prior predictive distribution may then result in incorrect decisions, in particular for θ values in the tails of the θ distribution. In future research, however, results from these Bayesian simulation methods should be compared. Another argument against using $f(\mathbf{x} \mid \hat{\theta})$ is that we have relatively long tests (between 70 and 140 items), thus $\hat{\theta}$ is estimated accurately.

The simplest way to obtain $f(\mathbf{x} \mid \hat{\theta})$ is to simulate item score patterns according to the IRT model and the selection algorithm used. An alternative is to use the empirical dataset at hand and select groups of examinees with approximately the same θ value and then determine *LB* and *UB* values for these groups of examinees. In this study we both simulated new data and used the empirical dataset to determine the bounds. A drawback of using the observed item score patterns may be that misfitting item score patterns could affect the values of the bounds. This effect, however, was considered to be small because the (realistic) assumption was made that almost all item score patterns would be in line with the underlying IRT model. Furthermore, we used the CUSUM procedure in this study only to illustrate and to explore its usefulness to detect unusual item score patterns. We will return to this topic in the discussion section.

Analysis

We analyzed the score patterns using the item ordering of presentation. Note that the method proposed by van Krimpen-Stoop and Meijer (2000) is based on the order of presentation. Bradlow et al. (1998) note that, using the order of presentation, warm-up outliers can be detected. Those are examinees who have trouble settling in or warming-up to the exam due to unfamiliarity or nervousness. As a result, the earliest

answers are more likely to be incorrect than the later answers. To increase the power, the lower boundary can be adapted by only taking the first a answers into consideration and by setting the upper bound equal to a value that can never be reached. For statistic (6) this would suggest setting UB greater than 1. After the first a answers, the lower bound is set to a value that never can be reached; that is, $LB < -1$. The choice of a may be based on a priori expertise knowledge or based on earlier observations. To detect item score patterns with many incorrect answers at the end of the test (due to, for example, tiredness), the item order can be reversed and the same methodology can be applied. Also, choosing $a_1 \leq k \leq a_2$ is possible (Bradlow et al., 1998). A limitation of this method is that to set these boundaries additional knowledge should be available. In our case it was difficult to predict how these boundaries should be chosen.

Relation between test length and misfit. Because the CAT we used had a varying test length, this enables us to investigate the relation between misfitting behavior and test length. In general it is expected that the proportion of misfitting item score patterns among the long tests may be an indication of misfitting behavior, in particular misfit at the start of the test. Of course long tests may also be the result of, for example, extreme θ values for which no "matching" item difficulties are available in the item bank. Therefore, long tests may not be identical with misfitting item score patterns, but misfitting item score patterns may result in long tests.

Results

Descriptive Statistics and Bounds

We analyzed the item score patterns of 1,392 examinees; 75.3% of the examinees obtained the minimum test length of 70 items, whereas to 11.1 % of the examinees the maximum test length of 140 items was administered. The mean of the final $\hat{\theta} = 1.83$ with $SD = .77$. The distribution of the final $\hat{\theta}$ is given in Figure 1. The mean item difficulty in the bank was 0.02 with a standard deviation of 1.04.

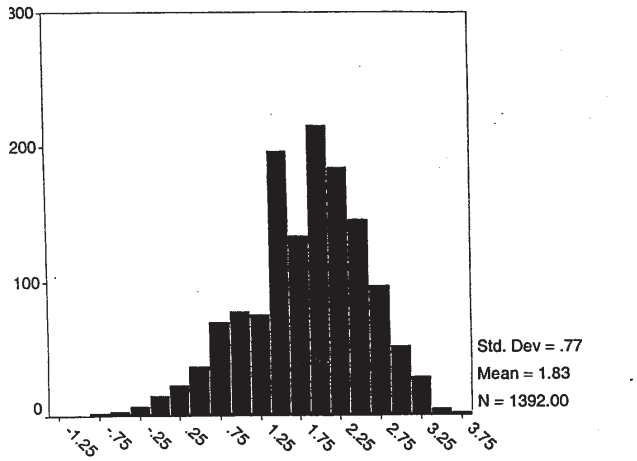

FIGURE 1. *Distribution of the final* $\hat{\theta}$

To analyze the item score patterns, we considered the item order as administered for each person. Because we need an upper and a lower bound to classify a pattern as fitting or misfitting, we determined these bounds by (1) considering all the 1,392 item score patterns in the sample and (2) conditioning on the θ level; this was done to investigate the effect of θ level on the height of the statistic. Note that the second strategy is in line with simulating item score patterns $f(\mathbf{x} \mid \hat{\theta})$, where $\hat{\theta}$ now was chosen as a class of values. To investigate the effect of conditioning on θ on the height of the LB and UB, we split the sample into three parts containing 33% of the lowest ($\theta < 1.536$), medium ($1.536 \leq \theta < 2.187$), and highest θ values ($\theta \geq 2.187$) respectively and determined the LB and UB in these subsamples.

Using all θ values to determine the lower and upper bound at a 5% level, we found $LB = -0.086$ and $UB = 0.109$. For $\theta < 1.536$ we found (0.113; –0.089); for $1.536 \leq \theta < 2.187$ we found (0.108; –0.084) and for $\theta \geq 2.187$ we found (0.108; –0.077). Thus, the bounds in these subsamples were almost the same as for the whole sample. These values are somewhat different from the values found in van Krimpen-Stoop and Meijer (2000). They found $LB = -0.13$ and $UB = 0.13$. The difference can probably be explained by the different item selection algorithm and the different distribution of the item difficulties. In their study, they used $b_i \sim N(-3, 3)$, whereas in this study the distribution of the item difficulties were normally distributed. To

investigate the influence of misfitting item score patterns, we also simulated 3,000 item score patterns based on the same distribution of $\hat{\theta}$ as discussed above and the same item bank using the Rasch model.

Similar bounds as discussed above were obtained. The response behavior of some of the persons with values below the *LB* and above the *UB* are presented in Figure 2.

Examples of misfitting item score patterns

For examinees #38, #262, and #488 the CUSUM crosses the *LB* and for examinees #312, #451, #503, and #683 the CUSUM crosses the *UB*. Let's consider these patterns in more detail. The CUSUM of examinee #38 with $\hat{\theta} = 0.136$ crosses the *LB* at the end of the test. Thus, at the end of the test many unexpected incorrect answers are given. Inspecting the pattern of item scores at the end of the test reveals that of the 17 last administered items (items #54–#70) 11 are answered incorrectly. Moreover, the mean *bs* of the incorrectly answered items equaled –0.560, which is unexpected given $\hat{\theta} = 0.136$. The same pattern occurs for person #262. For person #488 it is interesting that there are relatively many incorrect scores in the middle of the CAT. Inspecting the item scores of person #488 ($\hat{\theta} = 1.458$) revealed that of the first 13 administered items 11 were correctly answered, which resulted in a θ value around 2.0, but in the next 20 items, 14 are answered incorrectly with 7 items with *bs* between 0.36–1.590. At the second part of the CAT, 33 out of the 40 items were answered correctly, resulting in $\hat{\theta} = 1.458$. Note that the plot of the CUSUM gives information about the part of the test where misfitting behaviors occur, which may point at different types of deviant behavior. Many incorrect answers at the end of the CAT may be the result of tiredness, whereas many incorrect answers in the middle of the CAT may point at a temporary loss of concentration.

To illustrate the type of item score patterns with extreme positive CUSUM values, consider the examinees #451, #501, and #638. Person #451 ($\hat{\theta} = 2.62$) answered 6 out of the first 13 items incorrectly, which resulted in the administration of relatively easy items starting from item #14. Then, because of the many correct answers to the next items, this examinee obtains a relatively high θ value, which makes the incorrect answers to the first items unexpected. This same phenomena can be observed for person #503 ($\hat{\theta} = 2.29$). A different CUSUM pattern can be observed for examinee #638. This examinee with $\hat{\theta} = 0.745$ answers relatively many items correctly in the first part of the test, resulting after 28 items in a CUSUM value larger than the *UB*. This high occurance of correct item scores are unexpected, because in the second part of the CAT many easier items than the items in the first part are answered incorrectly. As a result the θ value levels off and becomes $\hat{\theta} = 0.745$. In particular the *correct* answers to items #6–#12 with *b* between 1.6 and 2.47 and #25–#32 with *b* between 1.3–1.25 are unexpected and result in a CUSUM value above the *UB*.

Because the test consists of different content areas, a possible explanation for this misfitting behavior is that the examinee masters some content areas better than others. Therefore, we determined the number of correct scores on the different content areas for the examinees with the CUSUM plots depicted in Figure 2. We did not, however, find a relation with content. Also, there was no relation between misfitting behavior and test length. Longer tests were administered to examinees with final $\hat{\theta}$ around the pass point.

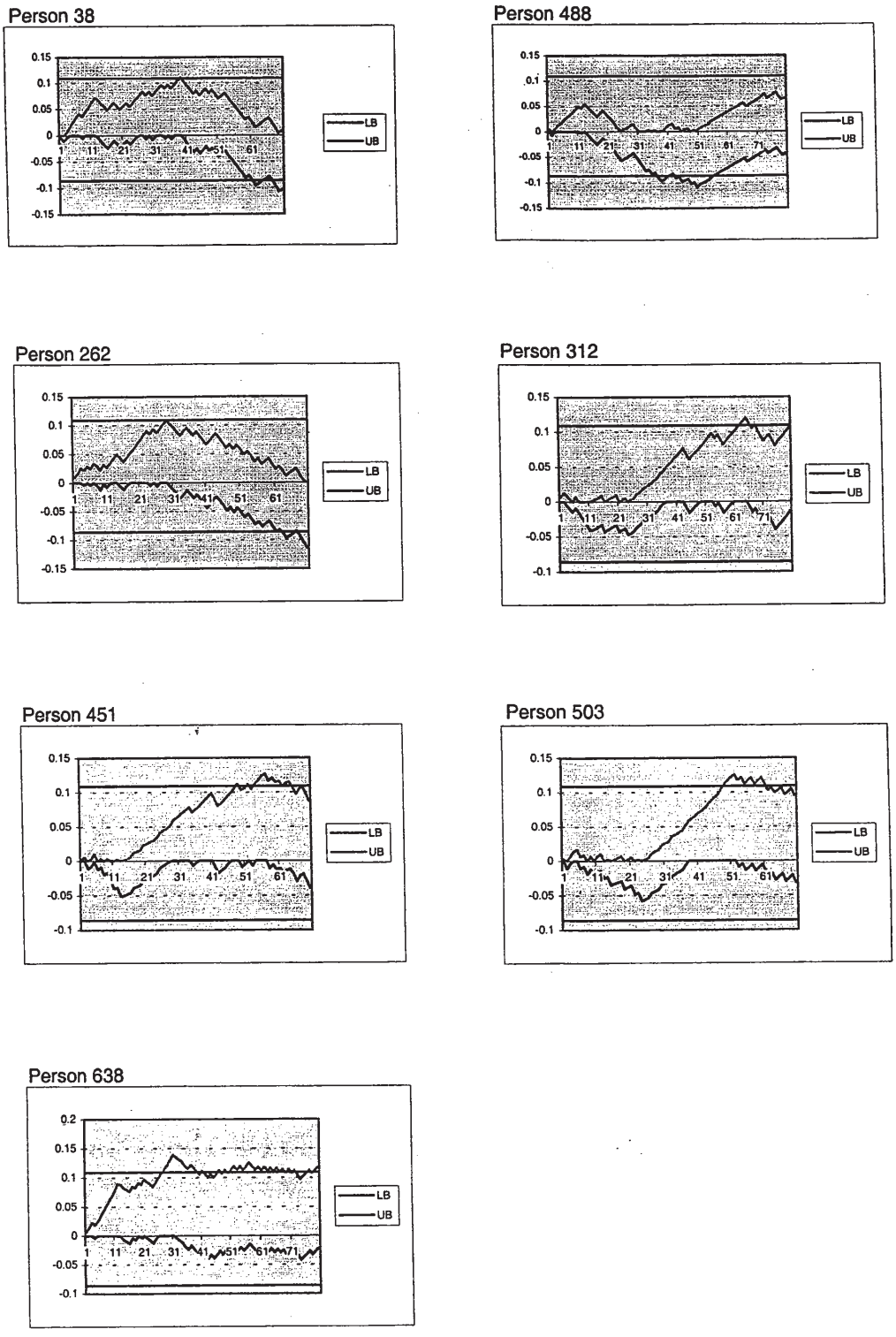

FIGURE 2. *Examples of the CUSUM for different examinees*

Discussion

In this study, we discussed three different methods to detect misfitting item score patterns in a CAT and applied one of these methods to illustrate the usefulness of this method to detect misfitting item score patterns. As the empirical analysis illustrated, item score patterns with values outside the bounds can be interpreted as having an item score pattern with unexpected responses. Note, however, that in performing an empirical analysis we do not know the *true* misfitting item score patterns, so we cannot report the detection rate. One of the advantages of using a CUSUM procedure as compared to paper-and-pencil person-fit statistics is that from the plots it is immediately clear where the aberrant behavior occurred in order of presentation as was illustrated above. This is a nice additional feature of the CUSUM procedure compared to general person-fit statistics. Model data fit can thus be investigated by local inspection of the CUSUM plot and this seems to be more useful than an overall statistic that only leads to the conclusion that an item score pattern does not fit the model. Moreover, by using the CUSUM procedure positive and negative strings can be distinguished.

Inspection of the CUSUM plots allowed us to distinguish different types of unexpected answering behavior. By using the order of presentation, it was possible not only to distinguish a warming-up effect, but also to detect examinees who may have become tired.

In this study we used the final $\hat{\theta}$ to calculate the CUSUM. An alternative may be to use the $\hat{\theta}$s that are estimated during the test administration. This would allow online information to be obtained concerning whether the examinee's responses are fitting the assumed IRT model. The problem with using the updated $\hat{\theta}$, however, is that it is based on few items (in particular at the first part of the test), which results in large standard errors. In practice this will not invalidate the use of the statistic, because, after an examinee has completed the test, researchers may want to investigate if the item score pattern is in agreement with the test model. If it is not, different actions can be taken depending on the type of test. If it is a low-stakes test used as a diagnostic tool in, for example, classroom assessment, valuable information about content may be obtained. To obtain that knowledge, the researcher can group the items according to their content and use the CUSUM to detect examinees that have difficulty in particular subject matters. Note that in this case additional information can be obtained and used by the teacher.

In high-stakes testing, the testing agency may use the CUSUM procedure to routinely check the data and compare the examinees outside the *LB* or *UB* across versions to ensure that the quality of the examination is the same across versions (using a fixed *LB* and *UB* for all versions). In addition, the possibility of preknowledge can be studied by checking the number of examinees falling outside the *UB*. Test takers may find the information useful. For example, consider person #38 who incorrectly answered many items at the end of the examination, items that according to the estimated ability and the responses earlier in the examination should have been answered correctly. This person might feel more confident in preparing for the next administration knowing that he/she had performed well early and being careful to guard against fatigue next time. This information is not available normally when the score reports contain only content subscores where, generally, the content is spread evenly across the examination.

References

Bradlow, F. T., Weiss, R. E., Cho, M. (1998). Bayesian identification of outliers in computerized adaptive testing. *Journal of the American Statistical Association, 93*, 910–919.

Drasgow, F., & Levine, M. V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement, 10*, 59-67.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67–86.

Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education, 9*, 47–64.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Kingsbury G. G., & Houser, R L. (1999). Developing computerized adaptive tests for school children. In: F. Drasgow & J. B. Olson-Buchanan, *Innovations in computerized assessment* (pp. 93–115). Mahwah, NJ: Lawrence Erlbaum.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics, 4*, 269–290.

McLeod, L. D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement, 23*, 147–160.

Meijer, R. R. (1998). Consistency of test behaviour and individual difference in precision of prediction. *Journal of Occupational and Organizational Psychology, 71*, 147–160.

Meijer, R. R., & Nering (1999). Computerized adaptive testing: Overview and Introduction. *Applied Psychological Measurement, 23*, 187–194.

Meijer, R. R, & Sijtsma, K. (1995). Detection of aberrant item score patterns: a review and new developments. *Applied Measurement in Education, 8*, 261–272.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: evaluating person fit. *Applied Psychological Measurement, 25*, 107–135.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55*, 75–106.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika, 41*, 100–115.

Rasch, G. (1960). *Probabilistic models for some intelligent and attainment tests*. Copenhagen: Nielsen & Lydiche.

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19*, 213–229.

Snijders, T. A B. (2001). Asymptotic distribution of person-fit statistics with estimated person parameters. *Psychometrika, 66*, 331–342.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (1999). Simulating the null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement*.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2000). Detecting person-misfit in adaptive testing using statistical process control techniques. In: W. J. van der Linden and C. A. W. Glas, *Computerized adaptive testing: Theory and practice*, (pp. 201-219). Boston: Kluwer Academic Publishers.

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

van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modem item response theory*. NY: Springer Verlag.

Wainer, H. (1990). *Computerized adaptive testing: A primer*. Hillsdale: Lawrence Erlbaum.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: Mesa Press.