ED 467 369 TM 034 295

AUTHOR	Meijer, Rob R.
TITLE	Outlier Detection in High-Stakes Certification Testing. Research Report.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
REPORT NO	RR-01-03
PUB DATE	2001-00-00
NOTE	30p.
AVAILABLE FROM	Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Adaptive Testing; *Certification; *Computer Assisted Testing; *High Stakes Tests; Item Response Theory; Licensing Examinations (Professions); Statistical Analysis
IDENTIFIERS	*Outliers; *Statistical Process Control

ABSTRACT

Recent developments of person-fit analysis in computerized adaptive testing (CAT) are discussed. Methods from statistical process control are presented 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 certification test were used. The item score patterns of 1,392 examinees were analyzed. Alternatives are discussed to generate norms so that bounds can be determined to classify an item score pattern as fitting or misfitting. Using bounds determined from a sample of a high-stakes certification test, the empirical analysis shows that the different types of misfit can be distinguished. Further applications using statistical process control methods to detect misfitting item score patterns are discussed. (Contains 2 tables, 3 figures, and 26 references.) (Author/SLD)

Outlier Detection in High-Stakes Certification Testing

Research Report
01-03

Rob R. Meijer

faculty of
EDUCATIONAL SCIENCE
AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

Outlier Detection in High Stakes Certification Testing

Rob R. Meijer

Abstract

Recent developments of person-fit analysis in computerized adaptive testing (CAT) are discussed. Methods from statistical process control are presented 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 certification test were used. Alternatives are discussed to generate norms so that bounds can be determined to classify an item score pattern as fitting or misfitting. Using bounds determined from a sample of a high-stakes certification test, the empirical analysis showed that different types of misfit can be distinguished. Further applications using statistical process control methods to detect misfitting item score patterns are discussed.

Index terms: computerized adaptive testing, item response theory, person-fit analysis.

Outlier detection in high-stakes certification testing

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

The development of CAT has resulted in more efficient 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 focus on the fit of an item score pattern to an IRT model in CAT, a research topic that has been under-exposed in the literature. For paper-and-pencil (P&P) tests many studies have focused on the fit of item-score patterns see Meijer and Sijtsma (1995; 2001) for a review. 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 item score pattern may be unlikely given the IRT model. Because in educational and psychological testing the main aim is to measure persons, the information that a person's item score pattern is unlikely given the model is valuable information. It may point at other mechanisms than the assumed interaction between the trait and item characteristics described by an IRT model. Therefore for both P&P tests and CAT identifying misfitting item score patterns is important, although the cause of misfit may be different for these 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 receive different tests.

Let us give two examples to illustrate the importance of investigating the fit of an item score pattern in a CAT and the mechanisms underlying aberrant response behavior. Kingsbury and Houser (1999) describe a 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 candidates mean achievement level. Some students may not take the test seriously and may guess the correct 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 instead of someone else's. 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 where a test is used as an instrument to assess the quality of a curriculum and where the results are not primarily used to evaluate an examinee.

As another example, consider high-stakes testing, where it is important that a test agency can 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 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

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

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

Most person-fit research used fit statistics designed to investigate the probability of an item score pattern under the null hypothesis of fitting response behavior. A general form for most person-fit statistics (Snijders, 1999) is

$$W = \sum_{i=1}^{k} [x_i - p_i(\theta)] w_i(\theta). \tag{2}$$

Most studies (e.g., Drasgow, Levine, & Williams, 1985; Levine and Rubin, 1979; Reise, 1995) have been conducted using some suitable function of the log-likelihood function

$$l = \sum_{i=1}^{k} \{x_i \ln p_i(\theta) + (1-x_i)\ln[1-p_i(\theta)]\}. \tag{3}$$

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 (e.g., Drasgow et al., 1985).

For relatively short P&P tests, the variance of this statistic is underestimated, that is, less than 1 (Nering, 1997), and corrections can be applied (e.g., Snijders, in press). Using statistics like 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

Insert Table 1 about here

on the person-fit statistic M proposed by Molenaar and Hoijtink (1990). This statistic is equivalent to l using the Rasch (1960) model and equals

$$M = -\sum b_i x_i. \tag{4}$$

Two sets of M-values are depicted, the first set (under M_1 in the Table) is based on the item difficulty values $(-2, -1, 0, 1, 2)$ and the second set (M_2) 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 #10 is the least plausible pattern. It can be seen that reducing the variance in the item difficulties also reduced the variance of M. In the extreme case when all items have the same item difficulty, all possible item score patterns have the same likelihood.

The situation with reduced variance in the item difficulties is relevant for person-fit in 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) who showed that in CAT the distributional characteristics of existing person-fit statistics like l are not in agreement with their theoretical distributions; the empirical type I errors were much smaller than the nominal type I errors.

Person fit in CAT

Few studies have proposed person-fit statistics using CAT. McLeod & Lewis (1999) proposed a statistic Z_c that is designed to detect item score patterns that result from memorization of the correct answers to an item. Before Z_c can be calculated the item bank is divided into three parts: easy items, items of medium difficulty, and difficult items. Let $\overline{Easy\left[p_i\left(\theta\right) - x_i\right]}$ denote the mean residual for the easy items and $\overline{Diff\left[p_i\left(\theta\right) - x_i\right]}$ the mean residual for the most difficult items in an administered CAT, and let k_{Easy} and k_{Diff} denote the number of easy and difficult items in the item bank, then Z_c is given by

$$Z_c = \frac{\overline{Easy\left[p_i\left(\theta\right) - x_i\right]} - \overline{Diff\left[p_i\left(\theta\right) - x_i\right]}}{\sqrt{\left\{\sum_{Easy}\left\{p_i\left(\theta\right)\left[1 - p_i\left(\theta\right)\right]\right\}/k_{Easy}^2\right\} + \left\{\sum_{Diff}\left\{p_i\left(\theta\right)\left[1 - p_i\left(\theta\right)\right]\right\}/k_{Diff}^2\right\}}} \tag{5}$$

Z_c is positive when an examinee answered the easy items incorrectly and the difficult items correctly which reflects misfitting answering behavior. 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. Also, not all administered items of a person are taken into account. This results in an incomplete picture of the fit of an examinee's item score pattern. Another drawback is that the statistic does not give information at what part of the test misfitting item scores occur.

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 modelled 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 a computer. Consequently, their responses to the first k_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 k_1 items were administered. Thus it was assumed that the final k_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

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

where $1/g$ is taken as the probability of a correct response during the subtest of k_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 formulate plausible alternative models. To how many items will the examinee guess the answers ? Will he/she completely guess without prior knowledge of the correct/incorrect answers ?

As an alternative to both methods discussed above, Bradlow, Weiss, and Cho (1998) and van Krimpen-Stoop and Meijer (2000; in press) proposed person-fit statistics based on the cumulative sum procedure (CUSUM, Page, 1954). Note that in CAT 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 indicate misfit or a bad bank. Sums of consecutive negative or positive residuals $[x_i - p_i(\theta)]$ can be investigated using a CUSUM. 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

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

Then, the sum of these $T_i s$ is accumulated as follows

$$\begin{aligned} C_i^+ &= \max\left[0, T_i + C_{i-1}^+\right], && (8)\\ C_i^- &= \min\left[0, T_i + C_{i-1}^-\right], \text{ and} && (9)\\ C_0^+ &= C_0^- = 0, && (10)\end{aligned}$$

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 $b_i \sim U(-3; 3)$. Furthermore assume that the items are multiple-choice items with 5 alternatives per item. 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. First, $\hat{\theta}$ is determined to investigate the fit of an item score pattern. 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 $\hat{\theta}$. Using $\hat{\theta}_{k-1}$ the fit can be investigated during test administration. Final $\hat{\theta}$ is more accurate and result are more stable (van Krimpen-Stoop & Meijer, 2000), therefore, final $\hat{\theta}$ is used in this example. Second, T is determined for each administered item, and third, based on the values of T and according to Equations

(8) through (10), C^+ and C^- are calculated for each administered item.

Consider an examinee responding the 20-item CAT generating the item score pattern given in Table 2. Final $\hat{\theta}$ for this examinee equalled $-.221$.Table 2 gives the values T, C^+, and C^- after the administration of each item. Consider the first

Insert Table 2 about here

three items. The first item score equals 0 and $p_i(\theta) = .411$, this results in $T_1 = -.021$ (Equation 7). Substituting this value in (8) results in $C_1^+ = 0$ and in (9) 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 $C_3^- = -.042 + .025 = -.017$. 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. This examinee may only take the test to get familiar with the test content. Randomly guessing the correct answer results in a probability of correctly answering the item of 0.20. In Table 3, the item score pattern for this examinee (final $\hat{\theta} = -3.5$) and T, C^+, and C^- are given. C^+ stays below .13, whereas at the 19th item the value of C^- becomes larger

Insert Table 3 about here

than $-.13$. As a result, the item score pattern is classified as *misfitting*. Note that although the final value C^- is smaller than $-.13$ we classify this pattern as misfitting because it crosses the LB which is unexpected compared to the score patterns in the norming sample. Note that the CUSUM takes the whole item score pattern into account and does not divide the item bank into different subsets of items as was done in the McLeod and Lewis (1999) study.

Van Krimpen-Stoop and Meijer (2000) used simulated data to investigate the power of the CUSUM procedure under different types of misfitting behavior. They simulated random response behavior, non-invariant ability (where two different θ values were used to generate the item responses for a simulee), and violations of local independence. They found detection rates of approximately .60 for guessing and between .22 and .72 for invariant ability depending on the θ level, for a 5% false positive rate. Detection of violations of local independence was poor (.10). Van Krimpen-Stoop and Meijer (2000) used simulated data to determine the power of the statistics. Below we will apply the CUSUM to empirical data.

Method

Data

We analyzed data from a high-stakes certification test to compare and investigate the usefulness of different person-fit methods. 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 topics. 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. There were 838 items in the item bank, and maximum likelihood estimation was used to estimate θ. Item exposure rate was controlled by a randomization algorithm set to choose an item within .5 logits of the targeted item difficulty.

Sampling distribution

A distribution of the statistics is needed to decide if an item score pattern is unlikely under the model. In CAT there are different possibilities for selecting the distribution $f(\mathrm{x})$ (Bradlow et al., 1998). The main question is what information we want to condition.

In this study we choose the simplest alternative, that is we used the distribution $f(\mathbf{x}|\hat{\theta})$ where we assumed that the item difficulties were known and fixed. Then, when we test 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). Because the test we analyzed has a variable length, we thus 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

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

or to sample from the prior predictive distribution

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

In general, sampling from the posterior or prior predictive distribution with, say, a normal prior may have the drawback that the fit of an item score pattern is compared with the fit of 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 (e.g., van Krimpen-Stoop & Meijer, 1999) . In future research, however, results from these Bayesian simulation methods can be compared. Another argument for using $f(\mathbf{x}|\hat{\theta})$ is that we have relatively long tests (between 70 and 140 items) and thus $\hat{\theta}$ is estimated accurately.

Item score patterns can be simulated according to the IRT model and the selection algorithm used to obtain $f(\mathbf{x}|\hat{\theta})$. 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 and not simulating according to the selection algorithm may be that misfitting item score patterns may effect the bound values. 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. 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 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 (7) this may equal $UB > 1$. After the first a answers the lower bound is set to a value that never can be reached $LB < -1$. 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, fatigue) 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 test length and misfit

Because the CAT has a varying test length (between 70 and 140 items), 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.

Results

Descriptive statistics and bounds

We analyzed the item score patterns of 1392 examinees; 75.3 % of the examinees obtained the minimum test length of 70 items, whereas 11.1% of the examinees the maximum test length of 140 items was administered. The mean of the final $\hat{\theta} = 1.83$ with a $SD = .77$. The distribution of final $\hat{\theta}$ is given in Figure 1. The mean item difficulty in the bank was 0.02 with a $SD = 1.04$.

Insert Figure 1 about here

The mean of the negative CUSUM values across $\hat{\theta}$ was -0.042 with a $SD = 0.021$ and the mean of the positive CUSUM values across $\hat{\theta}$ was 0.061 with a $SD = 0.032$. 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 1392 item score patterns in the sample and (2) conditioning on θ level, to investigate the effect of θ level on the distribution of the statistic. Note that the second strategy is in line with simulating item score patterns $f(x|\hat{\theta})$, where $\hat{\theta}$ is chosen as a class of values. To investigate the effect of conditioning on θ on the LB and UB we split the sample into three parts containing 33% of the lowest ($\hat{\theta} < 1.536$), medium ($1.536 \leq \hat{\theta} < 2.187$), and highest θ values ($\hat{\theta} \geq 2.187$), respectively and determined the LB and UB in these subsamples.

Using all $\hat{\theta}$ 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 $1.536 \leq \hat{\theta} < 2.187$ we found $(0.108; -0.084)$ and for $\hat{\theta} \geq 2.187$ 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 $b_i \sim U(-3; 3)$, 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 3000 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.

Examples of misfitting item score patterns

To illustrate the answer behavior of some persons with values below the LB and above the UB consider Figure 2. For examinees #38, #262, and #488 the CUSUM crosses the LB and for examinees #312, #451, #503, and #683 the

Insert Figure 2 about here

CUSUM crosses the UB. Let's consider these patterns in more detail. The CUSUM of examinee #38 with $\hat{\theta} = .136$ crosses the LB at the end of the test. Thus, at the end of the test many unexpected incorrect answers were given. Inspecting the pattern of item scores at the end of the test reveals that of the last 17 administered items (items #54-#70) 11 items were answered incorrectly. Moreover, the mean bs of the incorrectly answered items equalled $-.560$ which is unexpected given $\hat{\theta} = .136$. The same pattern occurs for person #262. Many incorrect answers at the end of the CAT may be the result of fatigue or guessing behavior as a result of lack of time to complete the test. However, the second explanation is unlikely because the examinee did not know how many items to expect. Note that the test length is variable (between 70 and 140 items) and depends on the accuracy by which a pass/fail decision can be made. 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 items were correctly answered which resulted in a θ value around 2.0, but then in the next 20 items 14 items were 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 $\hat{\theta} = 1.458$. Many incorrect answers in the middle of the CAT may point at a temporarily loss of concentration. Note, that the plot of the CUSUM gives information at what part in the test misfitting behaviors occurs which may point at different types of deviant behavior.

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 13 first items incorrect, 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

$\hat{\theta} = .745$ answered relatively many items correctly in the first part of the test, resulting in a CUSUM value larger than the UB after 28 items. This many 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 $\hat{\theta} = .745$. 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 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.

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 detect misfitting item score patterns. The empirical analysis illustrated that item score patterns with values outside the bounds can be interpreted as having item score pattern with unexpected responses. Note, however, that because in an empirical analysis we do not know the *true* misfitting item score patterns, 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 type of aberrant behavior is situated as illustrated above. This is a nice additional feature of the CUSUM procedure as 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, a CUSUM procedure allows positive and negative strings to be distinguished. The person response function discussed in Trabin and Weiss (1983) and Sijtsma and Meijer (2001) also allows for the inspection of local model violations, however, it is not based on the detection of strings of correct and incorrect item scores because it is formulated in the context of P&P testing.

By means of inspecting the CUSUM plots we could 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 became tired.

In this study we used the final $\hat{\theta}$ to calculate the CUSUM. An alternative is to use the $\hat{\theta}$ that is estimated during the testing process and by means of which on line information can be obtained if an examinee answers the items according to the 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 researchers may want to investigate whether the item score pattern is in agreement with the test model after an examinee has completed the test. 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 that can be used by the teacher. In high-stakes testing, the testing agency may use the CUSUM procedure to routinely check the data and compare the number of examinees outside the LB or UB across examinations to ensure that the quality of the examination is the same across examinations (using a fixed LB and UB across examinations). As a result of preknowledge of the items long strings of correct answers may occur more often resulting in a larger percentage of persons falling outside the UB. At the individual level, a CUSUM can be used to give the examinee insight in his or her answering behavior. For example, for a person failing the test such as person #38 in Figure 2 it is informative to know that many items were answered incorrectly at the end of the exam and answers in the first part of the test were answered correctly. This information may help the examinee to have confidence in his or her next examination and to know why he/she failed (for example due to nervousness at the second part of the test or fatigue). Note that when subtest scores on different content areas are reported this information is not obtained.

References

Bradlow, E.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, M. L. (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.

Nering, M. L. (1997). The distribution of indexes of person-fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21*, 115-127.

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.

Sijtsma, K. & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika*, *66*, 191-207.

Snijders, T. A. B. (in press). Asymptotic distribution of person-fit statistics with estimated person parameter. *Psychometrika*

Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D.J. Weiss (Ed.), *New horizons in testing*. New York: Academic Press.

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*, *23*, 327-345.

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 Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics*, *26*, 199-217.

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 Modern Item Response Theory*. NY: Springer Verlag.

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

Author note

This study received funding from the Law School Admission Council (LSAC). The author thanks the Council on Certification of Nurse Anesthetists and Computer Adaptive Technologies, Inc. for providing the data used in this research. The author also thanks Steve Reise and Michael Nering for their critical comments on an earlier version of this paper.

Table 1

M-values for different item score patterns

pattern						M_1	M_2
1	1	1	0	0	0	3	1.5
2	1	0	1	0	0	2	1
3	1	0	0	1	0	1	0.5
4	0	1	1	0	1	1	0.5
5	1	0	0	0	1	0	0
6	0	1	0	1	0	0	0
7	0	1	0	0	1	−1	−0.5
8	0	0	1	1	0	−1	−0.5
9	0	0	1	0	1	−2	−1
10	0	0	0	1	1	−3	−1.5

Table 2

Cusum Procedure for a Fitting Item Score Pattern

item	x	$p_i(\theta)$	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

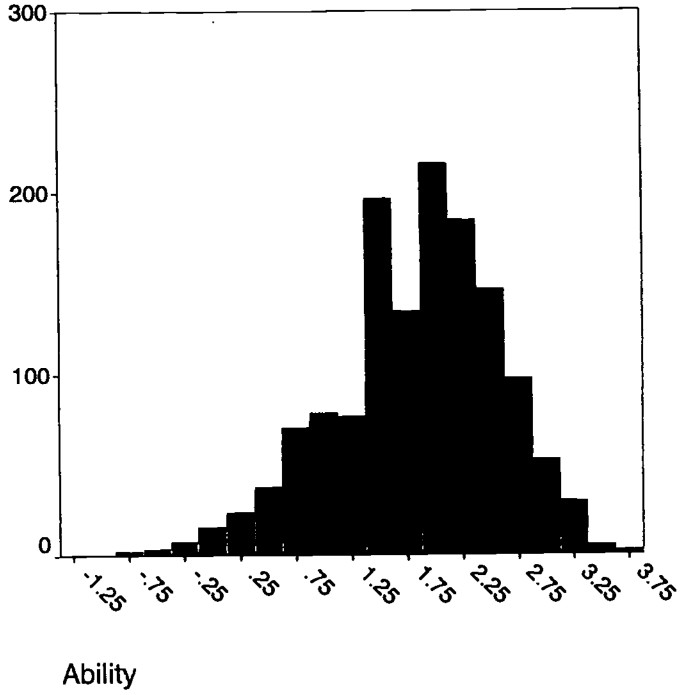
Table 3

CUSUM Procedure for a Misfitting (Guessing) Item Score Pattern

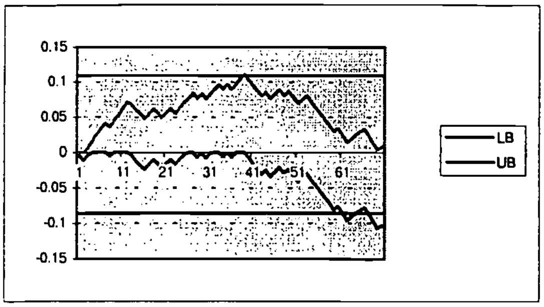
item	x	$p_i(\theta)$	T	C^+	C^-
1	0	.005	0	0	0
2	0	.006	0	0	-.001
3	1	.007	.050	.050	0
4	0	.009	0	.049	0
5	0	.012	-.001	.049	-.001
6	0	.026	-.001	.047	-.002
7	0	.060	-.003	.044	-.005
8	0	.070	-.003	.041	-.009
9	1	.134	.043	.084	0
10	0	.082	-.004	.080	-.004
11	0	.149	-.007	.073	-.012
12	0	.251	-.013	.060	-.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

Figure 1. Distribution of final $\hat{\theta}$

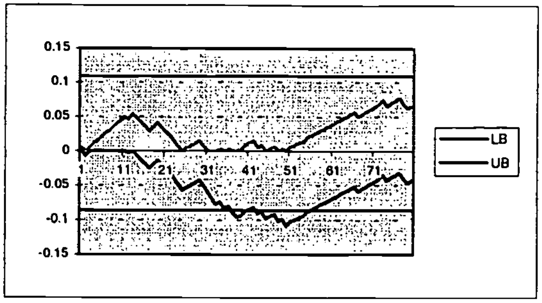
Figure 2. Examples of the CUSUM for different examinees


300
200
100
0
-1.25
-.75
-.25
.25
.75
1.25
1.75
2.25
2.75
3.25
3.75
Ability

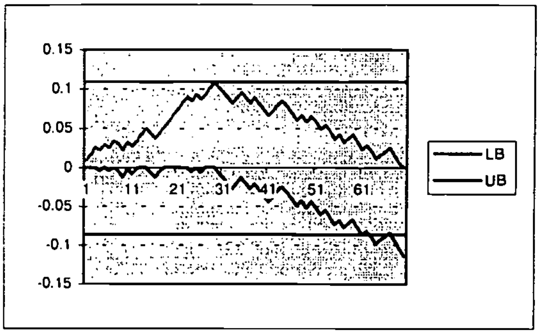
Person 38



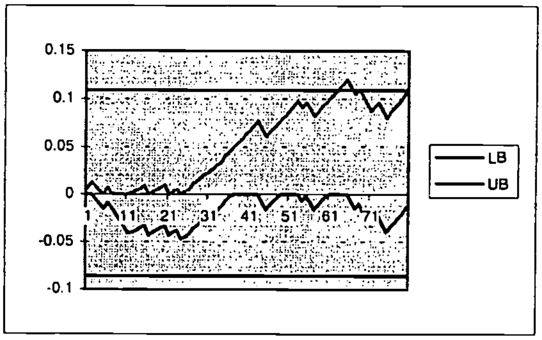
Person 488



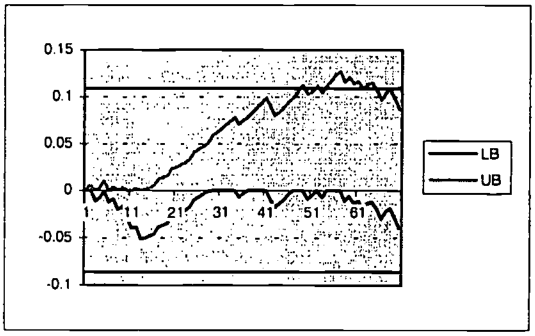
Person 262



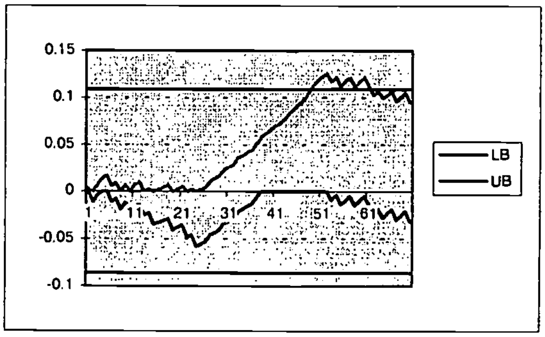
Person 312



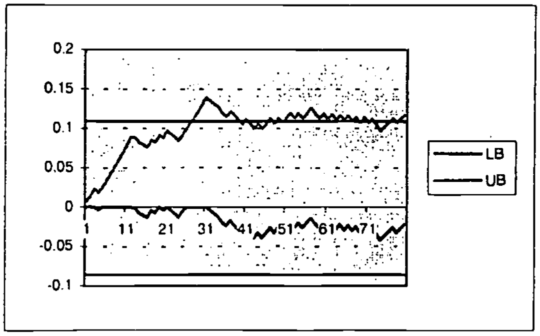
Person 451



Person 503



Person 638



BEST COPY AVAILABLE

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-01-03	R.R. Meijer, *Outlier Detection in High-Stakes Certification Testing*
RR-01-02	R.R. Meijer, *Diagnosing Item Score Patterns using IRT Based Person-Fit Statistics*
RR-01-01	W.J. van der Linden & H. Chang, *Implementing Content Constraints in Alpha-Stratified Adaptive testing Using a Shadow test Approach*
RR-00-11	B.P. Veldkamp & W.J. van der Linden, *Multidimensional Adaptive Testing with Constraints on Test Content*
RR-00-10	W.J. van der Linden, *A Test-Theoretic Approach to Observed-Score equating*
RR-00-09	W.J. van der Linden & E.M.L.A. van Krimpen-Stoop, *Using Response Times to Detect Aberrant Responses in Computerized Adaptive Testing*
RR-00-08	L. Chang & W.J. van der Linden & H.J. Vos, *A New Test-Centered Standard-Setting Method Based on Interdependent Evaluation of Item Alternatives*
RR-00-07	W.J. van der linden, *Optimal Stratification of Item Pools in a-Stratified Computerized Adaptive Testing*
RR-00-06	C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using a Multidimensional IRT Model and Bayesian Sequential Decision Theory*
RR-00-05	B.P. Veldkamp, *Modifications of the Branch-and-Bound Algorithm for Application in Constrained Adaptive Testing*
RR-00-04	B.P. Veldkamp, *Constrained Multidimensional Test Assembly*
RR-00-03	J.P. Fox & C.A.W. Glas, *Bayesian Modeling of Measurement Error in Predictor Variables using Item Response Theory*
RR-00-02	J.P. Fox, *Stochastic EM for Estimating the Parameters of a Multilevel IRT Model*
RR-00-01	E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Detection of Person Misfit in Computerized Adaptive Tests with Polytomous Items*
RR-99-08	W.J. van der Linden & J.E. Carlson, *Calculating Balanced Incomplete Block Designs for Educational Assessments*
RR-99-07	N.D. Verhelst & F. Kaftandjieva, *A Rational Method to Determine Cutoff Scores*
RR-99-06	G. van Engelenburg, *Statistical Analysis for the Solomon Four-Group Design*
RR-99-05	E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *CUSUM-Based Person-Fit Statistics for Adaptive Testing*
RR-99-04	H.J. Vos, *A Minimax Procedure in the Context of Sequential Mastery Testing*

RR-99-03 B.P. Veldkamp & W.J. van der Linden, *Designing Item Pools for Computerized Adaptive Testing*

RR-99-02 W.J. van der Linden, *Adaptive Testing with Equated Number-Correct Scoring*

RR-99-01 R.R. Meijer & K. Sijtsma, *A Review of Methods for Evaluating the Fit of Item Score Patterns on a Test*

RR-98-16 J.P. Fox & C.A.W. Glas, *Multi-level IRT with Measurement Error in the Predictor Variables*

RR-98-15 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using the Rasch Model and Bayesian Sequential Decision Theory*

RR-98-14 A.A. Béguin & C.A.W. Glas, *MCMC Estimation of Multidimensional IRT Models*

RR-98-13 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Person Fit based on Statistical Process Control in an AdaptiveTesting Environment*

RR-98-12 W.J. van der Linden, *Optimal Assembly of Tests with Item Sets*

RR-98-11 W.J. van der Linden, B.P. Veldkamp & L.M. Reese, *An Integer Programming Approach to Item Pool Design*

RR-98-10 W.J. van der Linden, *A Discussion of Some Methodological Issues in International Assessments*

...

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands.

faculty of

EDUCATIONAL SCIENCE AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands