A cat with personality and attitude: computerized adaptive testing of personality and attitude attributes with polytomous items

Hol, A.M.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Detecting aberrant responders to attitude questionnaires: Combining Guttman errors, testing times, and item proposition lengths

Abstract

A method is presented to detect aberrant responding. The method is based on the number of individual Guttman errors, the testing time, and the individual correlation of item proposition length and item response time. To demonstrate the method item response data were used that originated in an experiment that compared computerized conventional and computerized adaptive administration of a subscale of a motivation questionnaire. Results showed that most of the participants did not show aberrant response behavior, and apparently responded seriously to the questionnaire. A small minority had both small testing times, a large number of Guttman errors, and a negative individual correlation between item proposition length and testing time, all of which may be interpreted as indicators of aberrant response behavior. Generally, the results of the computerized conventional and the computerized adaptive test were similar. A few significant administration mode effects were found, however, these were of little consequence.
6.1 Introduction

The psychometric assessment of aberrant response behavior and person fit has mainly been concerned with the measurement of cognitive ability or achievement (e.g., Karabatsos, 2003; Meijer, 1994; Meijer & Sijtsma, 2001; van der Linden & van Krimpen Stoop, 2003). However, aberrant response behavior may also arise in the measurement of other psychological variables such as personality attributes and attitudes.

Karabatsos (2003) provided an overview of the factors that may lead to aberrant responses in ability or achievement testing. These include cheating (e.g., copying from another examinee), careless responding, lucky guessing, creative responding, and random responding. Apart from cheating and lucky guessing, these factors may also play a role in personality or attitude measurement. When nothing is at stake, candidates may get bored, and careless, which may result in creative or random responding. Additionally, other factors that arguably are specific to personality or attitude measurement may play a role, such as social desirable answering, acquiescence, and faking. For instance, job candidates may respond to personality or attitude questionnaires in such a way so as to increase the probability of getting the job.

Recent developments in computerized assessment and testing in psychology measurement offer the possibility of recording item response time accurately. In the context of ability measurement, item response times and aberrant response behavior have been studied in a simulation study (van der Linden & van Krimpen Stoop, 2003). However, to the knowledge of the present author, the combination of item response time measurement and person fit has yet to be considered in personality or attitude measurement.

The aim of the present paper is to present a method that combines the number of individual Guttman errors, individual testing times, and individual correlations between item proposition length and item response time to detect aberrant response behavior. To demonstrate the method, we analyzed data of students who as a part of the study were obliged to undergo a long testing session¹. Because of the length of the session and the obligatory nature of their participation, it is possible that some students did not participate seriously. This may have resulted in aberrant response behavior, such as responding very quickly, or responding in a way that results in a large number of personal Guttman errors. Negative individual

¹ Students who object to participating are assigned an alternative task. In practice, very few students choose the alternative assignment.
correlations between item proposition lengths and item response times may also be indicative of aberrant response behavior. If item propositions are simple to understand and to respond to, and participants read these propositions carefully, than a positive correlation between item proposition length and item response time can be expected. A negative individual correlation may indicate that the participant did not read propositions carefully.

It is likely that the combination of the number of Guttman errors with the two other indicators of aberrant response behavior is most useful in detecting such behavior, because a large Guttman error by itself may also be indicative of person misfit. If a participant shows aberrant response behavior on multiple indicators, his or her data may be considered unreliable.

To demonstrate the present method of detecting aberrant responders, we used item response data from a previous study (see chapter 4). In that study an experimental comparison was made between conventional computerized testing (CT) and computerized adaptive testing (CAT). Because the complete set of 24 items was administered adaptively in the computerized adaptive test (CAT), it was possible to compare this full scale CAT to the CT version. Therefore, the occurrence of Guttman errors and testing times could also be compared considering CT and CAT administration.

In the original study (see chapter 4) and in a second study (see chapter 3) that compared paper and pencil, CT and CAT administration, no effects of the type of test (i.e., CAT vs. CT) were observed. Therefore, substantial differences between CT and CAT were not expected with respect to Guttman errors and testing times.

6.2 Method

6.2.1 Participants
A total of 515 psychology freshmen participated in this study. 241 students participated in the CT condition and 274 participated in the CAT condition.

6.2.2 Materials
The test used in the present study was a 24 item motivation scale (MS) that measures attitudes towards academic education. The MS is part of the School Attitude Questionnaire (Vorst, 2000), which is used widely in high schools in the Netherlands. The items of the MS were adapted to the situation of academic education. Each motivation item consists of a proposition, and the participant is asked to indicate whether the proposition
is applicable to him- or her on a three-point Likert scale. This scale has three response options: that is the case, I don't know, and that is not the case. An example of a motivation item is: I usually study the subject matter of my study very well. Table 4.1 shows all 24 propositions. Of the 24 items, 12 items were stated positively with respect to the trait, and 12 items were stated negatively with respect to the trait.

The scale was constructed using the MSP program, which does Mokken scale analysis for polytomous items (Molenaar et al., 2000). This program is suited to select items and to assess the fit of Molenaar's (1997) extension of Mokken's (1971, 1997) nonparametric monotone homogeneity model (NMHM) to polytomous items. The calibration sample that was used to construct the scale consisted of the data of 1924 psychology freshmen of three previous years using a paper and pencil test. The original scale consisted of 48 items. The reduced scale of 24 items had a scale coefficient of $H=0.42$, which satisfies the recommended practical minimum lowerbound for a medium scale (Sijtsma & Molenaar, 2002). The Hi coefficients of individual items ranged from .30 to .54. The values of the H coefficients indicate that the scale predominantly measures one underlying attitude. Moreover, a simulation study (see chapter 4) showed that generation of item responses using the calibration sample item parameters of the MS given a more restrictive model nested under the NMHM resulted in a scale coefficient of $H=0.43$, which is close to the value of $H=0.42$ that was found for the real item response data of the MS. Cronbach's reliability coefficient of the present MS equalled $\alpha=0.92$.

### 6.2.2.1 Computerized versions of the MS

A computer program was written that administered the 24 motivation items one at a time. It was possible to change the response to an item while it was displayed on the screen. After continuing to the next item it was not possible to return to previous items. It was not possible to skip items without responding. The computer program also registered item response time, which was measured by computing the time between the moment of item display and the moment of item response.

The computer program assigned each of the participants randomly to one of two conditions, involving either CT or CAT administration of the scale.
CT Version

The CT administered items in the order in which they originally occurred in the paper version of the SQ. This order was identical to the order of items as shown in Table 4.1.

CAT Version

All 24 items were administered one at the time in an adaptive sequence (see chapter 4). Because all 24 items were administered in the CAT, the CAT version was completely comparable to the computerized version. The only difference between the two conditions was the order of presentation of the 24 items in the CAT version. This order varied from person to person.

6.2.3 Analysis

For each participant the number of Guttman errors, item response times, the total testing time (sum of the item response times), and the individual correlation between item proposition length and item response time were computed.

6.2.3.1 Computation of individual Guttman errors

In the present study, we employed Sijtsma and Molenaar’s (2002) definition of the Guttman error. The Guttman error is a very simple person fit statistic, which has proved to be a useful and simple alternative to more complex person fit measures (Meijer, 1994). The personal Guttman error of each participant was computed. Computation of the Guttman error requires the splitting of polytomous item scores into dichotomous item steps. The score on a polytomous item that consists of three ordered response categories (0, 1, and 2) is divided into two dichotomous item steps. The first item step is scored 0 if the response to the polytomous item is 0, and the first item step is scored 1 if the response to the polytomous item is 1 or 2. The second item step is scored 0 if the response to the polytomous item is 0 or 1, and the second item step is scored 1 if the score to the polytomous item is 2.

Computation of Guttman errors also requires knowledge of the order of the item steps. The order of item steps used here was the order that was found in the calibration sample of 1924 students of previous years. Guttman errors were computed by counting the times that the score on a more popular item step is 0, while the score on a less popular item step is 1.
6.2.3.2 The aberrant responders detection method

We detected aberrant responders by considering the combination of Guttman errors, testing times (sum of the item response times), and individual correlations between item proposition length and item response times.

Participants with fast response times may have larger Guttman errors, because fast response times may indicate that the participant did not read the proposition carefully, which in turn may result in an inconsistent response pattern. Fast reading of, and responding to propositions may very well result in inconsistent response patterns, as the set of items included propositions that were stated contra-indicative with respect to the latent trait. Students that had large Guttman errors despite a "normal" testing time may be indicative of model or person misfit due to other reasons.

To assess whether students had testing times and/or Guttman errors that can be considered as outliers, we used an outlier detection rule based on the median absolute deviation (Wilcox, 2003, p.77).

Because the item propositions were simple to understand and to respond to, careful reading should result in a positive correlation between item proposition length and item response time. This relationship was studied both on the aggregate level as on the individual level.

We studied the relation between item proposition length and the mean of the participants' item response times across items per condition (i.e., CT and CAT) on the aggregate level. The number of words of each of the items should correlate positively with the mean of item response times across the 24 items. Furthermore, we computed the individual correlations between item proposition length and item response time. Accordingly, we used three indices for the detection of aberrant responders: the number of Guttman errors, testing times, and individual correlations between item proposition length and item response time.

6.2.3.3 Administration mode effects

The effect of administration mode (CAT vs. CT) on a number of variables was investigated. Because both tests included all 24 items, any effect of mode should be due to the only difference between the tests, namely the order of the items.

The effect of mode on the number of Guttman errors was tested with the Mann-Whitney U test (Gibbons, 1985). A possible administration mode effect on the mean of total testing times was tested with a t-test. Furthermore, effects of mode on the correlation between item proposition length and item response times across the 24 items was tested using the above mentioned outlier detection rule.
response time across items, and on the correlation between testing times and the number of Guttman errors across participants were tested following a Fisher-z transformation. The difference between individual correlations between item proposition length and item response time in the CT and CAT condition were tested with the Mann-Whitney U test (Gibbons, 1985). All hypotheses were tested at the 5% significance level.

6.3 Results

6.3.1 The aberrant responders detection method

Figures 6.1a and 6.1b show the relationship between individual testing times and the number of individual Guttman errors for the CT and the CAT version, respectively. For the CAT, a significant negative correlation was found between testing time and the number of Guttman errors \( (r=-.234, \ p<.05) \). For the CT, this correlation was not significant \( (r=-.055, \ p=.40) \).

Figure 6.2a and Figure 6.2b show the relationship between the number of words in each proposition\(^2\) and the mean of item response times for each proposition for the CT and the CAT version, respectively. Both figures show that the result for one proposition is an outlier. For the CT version this proposition consists of 9 words, and it has a mean item response time of about 7 seconds. For the CAT version this proposition consists of 5 words,

![Figure 6.1](image)

**Figure 6.1.** Relationship between testing time and number of Guttman errors in the CT and the CAT group.

\(^2\) The numbers of words in Figure 6.2a and 6.2b correspond to the original Dutch propositions. The numbers of words in the translations in Table 4.1 are not equal to the numbers of words in the original Dutch propositions.
and it also has a mean item response time of about 7 seconds. In both versions this was the first proposition that was administered. Although the CAT administered items in adaptive order, the first proposition that was administered, was identical for all participants. In both versions, students probably had to adapt to the questionnaire format. The relationship between item proposition length and the mean of item response times was therefore absent for the first proposition. For the remaining 23 propositions there was a clear relationship between item proposition length and the mean of its item response times in both versions. For the 23 remaining propositions in the CT version, the correlation between item proposition length and the mean of its item response times was \( r=.876 \) (\( p<.05 \)). In the CAT version this correlation was \( r=.862 \) (\( p<.05 \)). Therefore, on group level there was generally a strong relationship between item proposition length and item response time, which indicates that most students participated seriously in this study. The relationship between item proposition length and item response time was also computed for individual participants. For the CT version, individual correlations between item proposition length and item response time had a median value of \( r=.528 \) and ranged between \( r=-.376 \) and \( r=.851 \). For the CAT version, individual correlations had a median of \( r=.408 \) and ranged between \( r=-.370 \) and \( r=.852 \). In the CT version 5 persons (2.1%) had a negative individual correlation, and in the CAT version 16 persons (5.8%) had a negative individual correlation.

The outlier detection rule, which is based on the median absolute deviation (Wilcox, 2003, p. 77), was used to identify students with extreme
testing times and extreme numbers of Guttman errors. Tables 6.2a and 6.2b contain the frequencies of the classification of the students in the different categories for the CT and the CAT version, respectively. In addition, Table 6.2a and 6.2b contain the number of positive and negative individual correlations between item proposition length and item response times.

For the CT version, the results in Table 6.2a show that two persons show aberrant response behavior on all of the three indicators. They had a large Guttman error, a small testing time and a negative individual correlation between item proposition length and item response time. Further examination of the data of these two persons showed that one of the two persons had many responses in the intermediate I don't know response category (14 responses). This may also have inflated his/her number of Guttman errors. For the other person the inflated Guttman error was probably caused by an inconsistent pattern of responses in all three categories.

For the CAT version, the results in Table 6.2b show that four persons have small testing times, large Guttman errors, and negative individual correlations between item proposition length and item response time. Detailed examination of the CAT item response data did not reveal any peculiarities that could explain the occurrence of large Guttman errors.

Generally, the results for both the CT and the CAT version showed that testing times and Guttman errors are related. This relationship is such that the majority of participants with small testing times have large Guttman errors, and a small portion of the participants with moderate testing times has large Guttman errors. Finally only one of the participants with very large testing times had a large Guttman error.

Of the 7 participants that had both a short testing time and a large Guttman error in the CT (3) and the CAT (4) version, 6 participants also had a negative individual correlation between item proposition length and item response time.

The data were explored for other possible phenomena concerning item response times, total testing times, and the number of Guttman errors. This exploration revealed that item response times and total testing times were unrelated with the total sum score on the MS.

Inspection of the item response data of the participants, who had large Guttman errors despite normal testing times, did not reveal any peculiarities. Since some students originated from other countries than the
Table 6.2a. Classification of testing times, number of Guttman errors, and number of positive/negative individual correlations between item proposition length and item response time for the CT version

Guttman errors

<table>
<thead>
<tr>
<th>Testing times</th>
<th>Positive</th>
<th>Negative</th>
<th>Positive</th>
<th>Negative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small outliers</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Normal</td>
<td>214</td>
<td>2</td>
<td>8</td>
<td>1</td>
<td>225</td>
</tr>
<tr>
<td>Large outliers</td>
<td>12</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Total</td>
<td>226</td>
<td>2</td>
<td>10</td>
<td>3</td>
<td>241</td>
</tr>
</tbody>
</table>

Table 6.2b. Classification of testing times, number of Guttman errors, and number of positive/negative individual correlations between item proposition length and item response time for the CAT version

Guttman errors

<table>
<thead>
<tr>
<th>Testing times</th>
<th>Positive</th>
<th>Negative</th>
<th>Positive</th>
<th>Negative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small outliers</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Normal</td>
<td>240</td>
<td>10</td>
<td>8</td>
<td>0</td>
<td>258</td>
</tr>
<tr>
<td>Large outliers</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>250</td>
<td>12</td>
<td>8</td>
<td>4</td>
<td>274</td>
</tr>
</tbody>
</table>
Netherlands, it is possible that language problems resulted in larger Guttman errors in these participants. However, most students, who had larger Guttman errors despite normal testing times, did not originate from other countries. Therefore, possible reasons for the larger Guttman errors in these students is that the NMHM model did not apply to these students, or that these students did not participate seriously despite their normal testing time.

6.3.2 Administration mode effects

A summary of the tests on administration mode effects is shown in Table 6.3. The Mann-Whitney U test applied to numbers of Guttman errors showed that the null hypothesis of equal medians between the CT and the CAT group was not rejected (z=-1.67, p=.095). Likewise the t-test applied to total testing times showed that the null hypothesis of equal means in the CT and the CAT group was not rejected (t=-1.01, p=.274).

The null hypothesis of equal correlations between item proposition length and mean of item response times in the CT and the CAT group was not rejected after Fisher-z transformation. The Mann-Whitney U test applied to the individual correlations between item proposition length and item response time indicated that the medians of the CT and the CAT group were not equal (z=-4.97, p<.05). The median of individual correlations was higher in the CT version (r=.528) than in the CAT version (r=.408).

The hypothesis of equal correlations between testing times and the number of Guttman errors in the CT (r=-.055) and the CAT group (r=-.234) was rejected after Fisher-z transformation (z=2.06, p<.05).

### Table 6.3. Administration mode effects of CAT versus CT administration

<table>
<thead>
<tr>
<th></th>
<th>CT</th>
<th>CAT</th>
<th>Statistic</th>
<th>Sign.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Guttman errors</td>
<td>70.9</td>
<td>80.7</td>
<td>z=-1.67</td>
<td>Ns</td>
</tr>
<tr>
<td>Testing times</td>
<td>80.7</td>
<td>82.7</td>
<td>t=-1.01</td>
<td>Ns</td>
</tr>
<tr>
<td>Correlation between item proposition length and mean of item response times</td>
<td>r=.876</td>
<td>r=.862</td>
<td>z=.64</td>
<td>Ns</td>
</tr>
<tr>
<td>Individual correlations between item proposition length and item response times</td>
<td>.528</td>
<td>.408</td>
<td>z=-4.97</td>
<td>p&lt;.05</td>
</tr>
<tr>
<td>Correlation between number of Guttman errors and testing times</td>
<td>r=-.055</td>
<td>r=-.234</td>
<td>z=2.06</td>
<td>p&lt;.05</td>
</tr>
</tbody>
</table>

Note: Sign. = Significance. Ns = Not significant.

1 Mean number of Guttman errors. 2 Mean of testing times (in seconds). 3 Median of individual correlations between item proposition length and item response time.

---

3 All students that reported to have entered the psychology curriculum with a non-Dutch high school diploma were considered as not originating from the Netherlands.
6.4 Conclusions

In the present paper, we proposed a method to detect aberrant response behavior in questionnaires. The method was demonstrated using CT and CAT data. No substantial differences were found concerning administration mode effects between CT and CAT data. Although other variables were studied, this finding is consistent with other research (see chapters 3 and 4), which found that CAT administration of polytomous attitude items could improve test efficiency, but did not lead to substantial administration mode effects.

The data of students were used to demonstrate the method. The fact that the participation of these students was obligatory\(^4\) and the test session lengthy could have resulted in aberrant test behavior and thus unreliable data. Since nothing was at stake, many students may have become bored, and have sought diversion in creative or random responding. However, the present results suggest that such behavior did not take place. The vast majority of students seemed to participate seriously, as was evident in the high positive correlation between item proposition length and the mean of item response times on the aggregate level in both the CT and the CAT version, which indicated that participants generally read propositions carefully. Furthermore, the raw sum scores of the motivation subscale correlated quite highly with the external measure mean of exam marks of the first three exams in the first year (CT: \(r=0.443\); CAT \(r=0.416\)).

Most of the participants, who had both a small testing time and a large Guttman error, also had a negative individual correlation between item proposition length and item response time. This suggests that a positive individual correlation between item response times and item proposition lengths provides an indication that students were serious and read propositions carefully. This relationship is plausible, because the most obvious explanation for the occurrence of a positive correlation between item proposition length and item response time is that participants read propositions, thereby positively influencing item response times by the item proposition length. When a participant does not read propositions well, chances are higher that the participant's item response times will not correlate positively with item proposition length.

In conclusion, the number of individual Guttman errors together with testing times, and individual correlations between item proposition length

\(^4\) Students who object to participating are assigned an alternative task. In practice, very few students choose the alternative assignment.
and item response time, provided a viable basis to detect aberrant response behavior. When aberrant response behavior is indicated by all three criteria, it is likely that participants were not responding seriously to the questionnaire.