Easy-to-Compute Response Times Based Statistics for Detecting Aberrant Behaviors of Test-takers

Zhen Li

eMetric, LLC

Nathan L. Wall

eMetric, LLC

August 8, 2022

Address all correspondence to: Zhen Li, Research Scientist, eMetric, LLC. 211 N Loop 1604 E, Suite 170, San Antonio, Texas 78232. Email: zli@emetric.net

Abstract

In this study, we develop an easy-to-compute statistic for detecting examinees' aberrant response times in large-scale computer-based assessments. Using this statistic, a response time is flagged as aberrant when it is longer or shorter than expected. The flagged response times are summarized to indicate examinees' abnormal test behaviors, such as pre-knowledge, rapid guessing and item memorization. A simulation study was conducted to evaluate this method's performance in various conditions. Results showed that the proposed statistic approximately followed a normal distribution in the null condition, performed equivalently well to van der Linden & Guo's (2008) Bayesian procedure in detecting aberrant response times, and reduced computational burden monumentally. A high-stake educational assessment was used to illustrate its application.

Key words: test security breach, response times, abnormal test behavior

Easy-to-Compute Response Times Based Statistics for Detecting Aberrant Behaviors of Test-takers

Introduction

A breach to test security may have serious implications for the psychometric integrity of the reported test scores and on the interpretations and consequences of those scores (Standards for Educational and Psychological Testing, 2014; p. 225). Statistical methods developed for test security purposes have become increasingly popular (Drasgow, Levine & Williams, 1985; Reise & Due, 1991; Impara et al., 2005; van der Linden & Guo, 2008; Marianti & et al., 2014; Li & Smith, 2015; Fox & Marianti, 2017). So far, statistical methods based on examinees' responses are the most studied, such as erasure analysis indices, answer copying indices and person fit indices. Applying statistical methods for test security purposes provides methods that are highly efficient but low cost, as they serve as a screening tool before more expensive investigations are conducted.

Compared to item responses, the study of response times is relatively new. According to recent research, response times can provide valuable information for improving test development, test security, as well as score use and interpretability. For example, Wise & Kong (2015) studied examinees' engagement during tests based on response time patterns. Fox & Marianti (2016) explore the relationship between response speed and accuracy. It is well known that traditional person fit indices examine the congruence between an item response pattern and a specified item response theory model (Reise &

Due, 1991). Similarly, aberrant test behaviors can be identified by comparing expected response times and observed response times. Recently, using latent variable modeling of response times, several methods have been developed for detecting aberrance in response time patterns (van der Linden & Guo, 2008; van der Linden, 2009). However, these methods have the shortcomings of being highly complex to implement and computationally ineffective. The purpose of this study is to develop easy-to-compute statistics for detecting aberrant response time patterns in large-scale assessments. The practical value of the three proposed statistics are illustrated and expanded in the paper. So far, no adequate approach has been developed to detect aberrant response time patterns during testing. Most of the current methods remain in the psychometric lab and are not widely used in practice. Either the field needs to develop new software to implement the more complex procedures based on response time models, or develop an easy-to-compute index, so that existing testing software can employ these methods without additional effort.

Data Forensic Methods and Response Time Modeling

Post-hoc data analysis for test security purposes has long existed. Early studies focused on the similarities between paired examinees' answers, indices of answer copying, and the likelihood of test-takers' responses with well-known test theory models. Nonetheless, test security breaches remain a significant issue in large-scale assessments since many existing statistical methods are only based on examinees' responses and a large number are highly complex. Recently, with the popularity of computer-based assessments, more information can be captured during the testing process. Therefore, new statistics have been proposed for this purpose.

Early studies focused on developing statistics used for detecting aberrant item response patterns, which are commonly known as person-fit indices (Reise & Due, 1991). The shift to computer-based testing allows examinees' response times to be easily recorded, however, and the response time data can be used for test security analyses. van der Linden & Guo (2008) proposed a Bayesian procedure to identify aberrant response time patterns, specifically those indicating pre-knowledge and item memorization in adaptive testing. Their method was based on a hierarchical latent response time model. In their first step, latent variables and item parameters are estimated based on a set of real data using Monte-Carlo Markov-Chain (MCMC). In the second step, the posterior distribution of each test-taker's response time is calculated. In the third step, the observed response time for each test-taker on each item is compared with the posterior distribution. A *p* value is computed, assuming that the posterior distribution is log-normal.

Qian et al. (2016) applied the Bayesian procedure to detect item pre-knowledge and potentially compromised items in two computer-based large-scale licensure examinations. The results indicated this procedure was helpful in monitoring aberrant examinee behaviors, as well as enhancing future item writing. However, this is the only publication that could be found applying the Bayesian procedure.

Marianti & et al. (2014) developed another set of statistics for detecting aberrant test behaviors based on the lognormal response time model. The statistics were derived from the well-known person fit statistic l_z (Reise & Due, 1991). The log likelihoods of the response time patterns were used to evaluate the fit of a response time pattern to a specified model. Furthermore, Fox & Marianti (2017) developed a person fit index with a hierarchical response time model, taking into consideration both response time and response accuracy. Additionally, Wang & Gong (2015) proposed a hierarchical mixture response time model for detecting examinees' engagement during testing, however this may be harmful to test validity. All these methods have their own merits in different ways. Nonetheless, the complexity of the methods based on latent variable modeling constrain them from being widely used in practice. Additionally, these methods are often computationally demanding.

Response Time Models

Using response time to explore and interpret test-takers' testing behaviors and outcomes is not necessarily a new approach. Thissen (1983) developed an extended item response theory model, taking into account person speed and ability simultaneously. Wang & Hansen (2005) developed a four-parameter logistic response time model, incorporating response time information to predict the probability of answering an item correctly. With these models, however, the aberrance of response time patterns is not known.

One of the popular response time models is van der Linden's (2006) lognormal model. The model assumes that test-takers' speed remains constant during testing, and that examinees answer each item independently. It contains three latent variables: τ_j represents the speed of test taker j, β_i is the time intensity of item i, and α_i is a discrimination parameter. This model is often presented as a lognormal density for the distribution of test-takers' response times (RT_{ij}):

$$f\left(R\mathrm{T}_{ij}; \tau_j, \alpha_i, \beta_i\right) = \frac{\alpha_i}{RT_{ij}\sqrt{2\pi}} e^{\{-\frac{1}{2}[\alpha_i(logRT_{ij}-(\beta_i-\tau_j))]^2\}}. \tag{1}$$

The advantages of this model include that, when it is combined with an item response model, a hierarchical response time model is formed. Not only could we study the relationship among latent variables, the information from response times might also be used for improving item calibration and test scoring.

van der Linden and other researchers have applied the above-mentioned response time model in many aspects. For example, van der Linden & Guo (2008) applied the hierarchical lognormal response time model for detecting aberrant test-takers' aberrant test behaviors in computer adaptive testing (CAT); van der Linden (2009) proposed another bivariate lognormal response time model for the detection of collusion between test-takers. van der Linden (2008) used this model to improve the accuracy of item selection in a CAT design.

As mentioned above, the idea of detecting aberrant test behaviors is not new. Any differences between the expected response pattern and the observed response pattern or the expected response time pattern and the observed response time pattern could imply an aberrant test behavior. The difficulty comes from how we detect the aberrances efficiently and accurately. Although van der Linden and Guo's (2008) Bayesian estimation can provide relative accurate latent variable estimates, the estimation process is computationally demanding. In the following section a simplified approach is proposed for detecting test-takers' aberrant behavior. Not only is this approach easy to implement in any testing software, it requires much less computation.

The Proposed Statistics

Basic Summative Statistics and the Indicator of Aberrance

Suppose there are *N* examinees and *n* items. The response time of examinee *j* to item *i* is indicated by RT_{ij}. First, the average response time to item *i* is indicated by $\overline{RT_{\iota.}}$, which is

$$\overline{RT_{\iota.}} = \frac{\sum_1^N RT_{ij}}{N}. \tag{2}$$

Furthermore, the total test time of examinee *j* is indicated by $RT_j = \sum_1^n RT_{ij}$. The average total test time among *N* examinees is:

$$\overline{RT_J} = \frac{\sum_1^N RT_j}{N} = \sum_1^n \overline{RT_{\iota.}}, \tag{3}$$

which is also equal to the sum of the average response time on each item. Thereby, the expected response time for examinee *j* on item *i* could simply be the average response time across all persons and all items: $\widehat{RT_{ij}} = {\overline{RT_j}}/{n}$.

An aberrance indicator is defined as the difference between the observed response time RT_{ij} and the expected value $\widehat{RT_{ij}}$:

$$Aberrance = (RT_{ij} - \widehat{RT_{ij}}). \tag{4}$$

Statistic I: Standardized Aberrance

Assuming that examinees' response time to each item follows a normal distribution, the simplest statistic for detecting aberrant response times can be obtained by standardizing the index of aberrance in equation 4:

$$Z_0 = \frac{RT_{ij} - \widehat{RT_{ij}}}{S.D.(RT_{i.})}. \tag{5}$$

Z_0 will be used as a baseline for index comparison. It is conjectured that it won't work well, as response times don't often follow a normal distribution.

Statistic II: Standardized the Aberrance using Logarithm of Response Times

It is commonly shown that students' response times follow a log-normal distribution. Therefore, we take the logarithm of the response time matrix, and compute a standardized residual index like this:

$$Z_l = \frac{\log(RT_{ij}) - \log(\widehat{RT_{ij}})}{S.D.(\log(RT_{i.}))}. \tag{6}$$

Notice that both Z_0 and Z_l have a strong assumption that all test-takers have an equal speed during testing. This assumption can often be violated in the real world. Following the idea of a standardized response time, a relative standardized response time is proposed. This new statistic considers person speed in the computation.

Statistic III: Standardized Logarithm Aberrance with Person Speed

Van der Linden (2006, 2009, 2011) defined person speed as latent variables in his lognormal RT model. As we know, person speed is defined as the item's time loading divided by response times on the item: $\tau_j = \frac{\beta_i}{\log{(RT_{ij})}}$, where τ_j is assumed to be constant across all items.

The definition of person speed used here follows the same structure. However, instead of using the latent variables, simple summative statistics are used. Assuming that person speed is a constant parameter across all items. Examinees' speed is quantified as the average total test time divided by the total test time of examinee *j*:

$$Speed_j = \frac{\overline{RT_J}}{RT_j}. \tag{7}$$

The longer the time an examinee takes for a test, the lower the speed is. With speed computed, the expected response time of examinee *j* to item *i* is obtained, which is the average response time on item *i* divided by person speed:

$$\widehat{RT_{\iota J}} = \frac{\overline{RT_{\iota.}}}{Speed_j}. \tag{8}$$

Similarly, when the response times are transferred onto the logarithmic scale, person speed can be calculated by $\overline{\log(RT_J)} - \log(RT_j)$. Furthermore, considering random errors reflected in the variances of response times on each item, person speed is calculated by the weighted average response time residual:

$$speed_j = \frac{\sum_{i=1}^{n} \frac{(\overline{\log(RT_\iota)} - \log(RT_{ij}))}{var(logRT_i)}}{\sum_{i=1}^{n} \frac{1}{var(logRT_i)}}. \tag{9}$$

The expected value of $\log(RT_{ij})$ after adjusting for person speed is $(\overline{\log(RT_\iota)} - speed_j)$.

The variance of $\log(RT_{ij})$ is the total variance minus the group (person) variance:

$$var(logRT_i) - var(speed_j). \tag{10}$$

Thereby, the standardized value of aberrance taking into consideration person speed variance is as follows:

$$Z_s = \frac{\log(RT_{ij}) - (\overline{\log(RT_\iota)} - speed_j)}{\sqrt{var(logRT_i) - var(speed_j)}}. \tag{11}$$

In the following two sections, we use a simulation study and an empirical study to examine the performance of these three proposed statistics.

Simulation Study

A simulation study was conducted to examine the performance of the proposed statistics in various conditions. In the simulation study, the proposed three statistics based on Z-score method were compared with two statistics based on van der Linden & Guo's latent variable modeling method. Furthermore, four factors, including proportion of compromised items, sample size, test length, as well as the population aberrance rates, might have an influence on the performance of these statistics. The effects of the four factors were examined in this study.

Simulation Design

Four factors were considered in the simulation study. Sample size and number of items in the test were considered because both are critical features of any assessment. A pilot study showed that the proposed statistic performed equivalently well in various conditions of sample sizes (500, 1000, 10000). Therefore, in the current simulation study, only the small sample size (*N*=500) condition was considered. Additionally, the aberrance rates of items (proportion of compromised items), and the aberrance rates in the sample (proportion of test-takers who have aberrant test behaviors) were considered based on previous research (Marianti & et al., 2014). The details of these factors are listed in Table 1.

Data Generation

1) Generating observed response times

Response times were generated based on a lognormal model (van der Linden, 2006):

$$\log(RT_{ij}) = \beta_i - \tau_j + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N(0, 1/\alpha_i^2) \tag{12}$$

where β_i, τ_j, and α_i were obtained from empirical data analysis. The longest response time for an item is constrained to be 20 minutes, and the shortest response time on one item is constrained to be 1 second, according to students' observed response times for multiple choice items in state assessments. Outliers, e.g., an extremely long response time or short response time, were not a factor of interest, therefore were not simulated in the current study.

2) Generating aberrant response times

Aberrant RTs were generated by three steps. First, a proportion of items were chosen according to the factor ARI in the simulation design. Second, a proportion of examinees were randomly selected from the sample according to the factor ARS in the simulation design. Third, with true person parameters and item parameters, the lognormal distribution of response times for the selected person on the chosen item was known. An aberrant RT is generated by taking a value of the cut points which are located at three standard deviations from the mean (half negative and half positive).

Statistics of Interest

The proposed three easy-to-compute statistics (Z_0, Z_l, Z_s) are of major interest in this study. In addition, two statistics based on van der Linden's (2006) lognormal response time model are computed for comparison.

After fitting the generated response time matrix with the lognormal model, parameter estimates $\widehat{\beta_\iota}$, $\widehat{\alpha_\iota}$ and $\widehat{\tau_J}$ can be obtained. Assuming that these parameter estimates were the true parameters, the expected response time for examinee *j* on item *i* is $(\widehat{\beta_\iota} - \widehat{\tau_J})$, and the variance of response times on this item as $({}^{1}/{(\widehat{\alpha_\iota})^2})$. A standardized residual index for detecting aberrance will be the fourth statistical index:

$$Z_m = \frac{\log(RT_{ij}) - (\hat{\beta}_i - \hat{\tau}_j)}{1/\hat{\alpha}_i}. \quad (13)$$

The reason for introducing Z_m is that it has the same structure as the proposed statistic Z_s. The only difference is that Z_m uses parameters estimated from latent variable modeling for persons' speed and items' time loadings.

The other statistic is a posterior predictive checking method, taking into consideration the posterior distribution of the person speed parameters. This method was first introduced by van der Linden & Guo (2008), known as a Bayesian procedure. Based on the posterior distribution of response time of person *j* on item *i*, the *p* values and standardized residuals were calculated by comparing the observed response time to this posterior distribution.

These two indices are comparable to the proposed easy-to-compute statistics above, as both of them follow a standard normal distribution. The difference lies in whether the estimated parameters from latent variable modeling or the observed response times are used directly to approximate the expected response times.

Evaluation Criteria

To determine whether a response time pattern is aberrant, the five statistics are calculated for each observed response time. RTs with extreme values will be flagged if the statistic is higher than 1.96 or lower than -1.96. After this process, each examinee will have or not have several flagged RTs. The proportion of flagged RTs of an examinee is computed. This number is then compared with a designated cut (varies with total number of items, see attachment). The power of the statistics is compared with respect to the following criteria:

Probability of detection: proportion of examinees flagged with aberrant RT patterns, among the actual number of examinees with true aberrant RT patterns (as designed in data generation). It is often regarded as a sensitivity index in statistics:

$$Detection\ rate = \frac{N * P_{true\ positive}}{N * (P_{true\ positive} + P_{false\ negative})}, \tag{14}$$

Precision of detection: Proportion of examinees with true aberrant RT patterns, among the number of examinees flagged as having an aberrant RT pattern. The equation is as following:

$$Precision\ of\ Detection = \frac{N * P_{true\ positive}}{N * (P_{true\ positive} + P_{false\ positive})}. \tag{15}$$

Results of Simulation Study

1) Type I error

In the null conditions, the asymptotic distributions of the proposed statistics are examined. As the statistics follow a standard normal distribution, when the α level was set to be 0.025 on each side, the empirical rejection rates at both sides should be close to 0.025.

In Table 2, the empirical rejection rates for Z_l and Z_s approximate 0.025 on both the left side ($\alpha = -0.025$) and the right side ($\alpha = 0.025$). However, the empirical rejection rates for Z_0 were not close to 0.025 on both sides. The large value of empirical rejection rates on the right side shows that the distribution of Z_0 was positively skewed.

In addition, flagging rules for individuals were used to check how likely an individual with no aberrance will be falsely flagged by the statistics. Results in Table 2 show that Z_s has zero probability of randomly flagging any examinee when no aberrant response time pattern exists. On the contrary, both Z_0 and Z_l flagged 6%-11% examinees incorrectly by chance.

2) Power

The power of the proposed statistics was tested in various alternative conditions, where aberrant response times were generated. The probability of detection (P.d.) and precision of detection (P) of statistics Z_0, Z_l and Z_s were compared with van der Linden & Guo's (2008) Bayesian procedure and Z_m.

From Table 3, it is obvious that Z_s performed better than Z_0 and Z_l. It had a higher precision of detection in all the conditions, and had a higher probability of detection in most conditions. When the number of items were small (n=20), and the proportion of compromised items was low (10%), which meant that only 2 items were compromised in the test, all statistics had very low (<= 0.08) probability of detection. However, the precision of detection for Z_s was much higher than those of Z_0 and Z_l. As the number of items increased and the proportion of compromised items increased, both the detection rates and precision of detection for Z_s grew. For example, when the number of items was 70, aberrance rates in the sample were 0.05, and the proportion of compromised items was 25%, the detection rate grew to 97%, and the detection precision grew to 98%. Furthermore, compared to the two approaches based on latent variable modeling (van der Linden and Guo's Bayesian procedures and Z_m), Z_s performed equivalently well in all the conditions we tested. The detection rates and precision of detection of the proposed new statistic were close to those of van der Linden and Guo's (2008) Bayesian procedure. In the following section, some factors are discussed that exert an influence on the performance of the proposed statistic.

3) The Influence of Factors

To better illustrate how each of the statistics' performance were influenced by the three factors: the proportion of compromised items, test length and the proportion of examinees with aberrant response times, the following figures are displayed.

Figure 1 shows the influence of aberrance rates of items on the probability of detection and precision when the number of items was increased from 20 to 70. The aberrance rate in the sample was fixed to 10%. When the percentage of compromised items was 10%, all five indices had low probability of detection. However, when the percentage of compromised items increased from 10% to 25%, three of the indices' probability of detecting aberrant response time patterns improved adequately for the short test (n=20), and improved rapidly for longer tests (n=40 and n=70). Z_m had the highest probability of detection in the tested conditions, while Z_s and the Bayesian procedure followed very closely. As a comparison, Z_0 and Z_l didn't improve significantly when the percentage of compromised items increased.

The bottom three plots in Figure 1 show that, Z_s, the Bayesian procedure and Z_m also performed similarly with respect to precision of detection in various conditions. When the aberrance rates of items increased to 25%, the precision of detection improved rapidly for the short test (n=20). For longer tests, the precision of these three detection indices was high even if the aberrance rate of items was low (10%), therefore the increased rates were small. When the aberrance rate of items reached 25%, no matter how many items there were, their precision of detection was close to 1. On the contrary, Z_0 and Z_l had very low precision of detection across all condition levels, with slight improvement when the aberrance rates of items increased to 25%.

Figure 2 shows the influence of aberrance rates of sample on the probability and precision of detection. The top plots in Figure 2 show that when the aberrance rates of sample increased from 5% to 10%, the three statistics with high detection rates (Z_s, Bayesian procedure, and Z_m) had lower probabilities of detecting aberrant response time patterns. The other two statistics (Z_0 and Z_l) had equivalently low probability of detection in both conditions. On the contrary, the bottom three plots of Figure 2 show that with the increased aberrance rate in the sample, all statistics' precision of detection improved slightly.

Both Figure 1 and Figure 2 show the influence of test length on the performance of the statistics: with the increased total number of items, the probability and precision of detection improved for all statistics. In particular, the precision of detection increased to almost 1 with a test length of 70, even if the percentages of compromised items and the percentages of aberrant examinees were relatively small.

4) Summary

A simulation study was carried out to evaluate the performance of the proposed statistics compared to an existing Bayesian procedure (van der Linden & Guo, 2008). In the null condition, two of the three new statistics, Z_l and Z_s, approximately follow a standard normal distribution. The simplest statistic Z_0 turns out to be extremely skewed to the right. Specifically, the left-tail p values were close to 0, while the right-tail p values approximated 0.05.

A most interesting finding was that, compared with the two procedures based on latent variable modeling, Z_s performed equivalently well in detecting response time aberrance. Among the proposed easy-to-compute statistics, Z_s performed much better than Z_0 and Z_l. Not only did it have higher precision of detection in all the simulation conditions, it also had a higher probability of detection in most conditions.

Moreover, several factors, especially the aberrance rates in the sample and the proportion of compromised items, exerted a significant influence on the performance of the proposed statistics. The more compromised items the test had, the less the examinees with aberrant response time patterns, the more likely that an aberrant response time pattern could be detected (high probability of detection). However, the precision of detection increased with both the number of compromised items and the proportion of examinees with aberrant response time patterns.

It was also found that the compromised items will always have more aberrant RTs than the other non-compromised items. In other words, based on the aberrance flagging of RTs, the compromised items were flagged correctly.

Empirical Study

Examinees' response times in a computer-based Math test consisting of 58 items were analyzed to illustrate the application of the proposed statistics. This test was part of a K-12 state testing program from the 2016-17 school year. The sample included 6,827 Grade 6 students. Demographically, the sample was diverse. Results from the real data analysis are discussed below.

We found that examinees' response times to most items approximately follow a lognormal distribution. One item was removed from analysis as its response time did not fit the lognormal distribution. The QQ-plot in Figure 3 shows that the total response time on the remaining 57 items was well approximated by a lognormal distribution, even though the slightly uplifted right tail indicated this distribution was a little heavy tailed.

Next, we computed the proposed statistics, as well as statistics based on Guo and van der Linden (2008)'s Bayesian procedure. Only a small proportion of aberrant response time patterns were detected (Table 4). With a close look at these aberrant response time patterns, possible reasons of the detected aberrance include speededness towards the end, time management strategies, and rapid guessing. For example, one examinee spent 1-5 seconds through all 57 items, which indicates a rapid guessing behavior. It appeared that no test-taker was engaged in any potential cheating behavior during testing.

Results in Table 4 show that about 5% of the examinees' response times were flagged by our statistic of interest (Z_{s}). According to the property of standard normal distribution, this probability is very close to the significance level 0.05. This finding indicates that no significant aberrance was found in the empirical data set. In addition, it was noticed that more positive response times were flagged than the negative response times by Z_0 statistic, which further proved a positively skewed distribution of response times. For Z_{l}, Z_s, and the Bayesian procedure, the difference between positive and negative flagging decreased to a small amount. Moreover, the extent to which two statistics flag the same examinee's response time pattern was checked.

In Table 5, we found that, among the flagged individuals, more than 88% were flagged simultaneously by Z_s and the Bayesian procedure, even if none of the examinees were identified with cheating behaviors. It is very likely that, when real aberrant response times exist, Z_s could be used to detect the aberrance as accurately as the Bayesian procedure, which was also proved in the simulation study.

Additionally, the percentages of flagged examinees by each statistic in each school were computed. Schools were flagged when the percentage of aberrant examinees was higher than the state-level percentage (bottom row in Table 4). Results show that 23%, 29%, 26%, 27% and 26% schools were flagged by Index $Z_{\mathbf{0}}$, Z_l, Z_s, Bayesian procedure, and Z_m respectively. Specifically, 96% of 596 schools in the state were simultaneously flagged by both Z_s and the Bayesian procedure. Meanwhile, only 63% of schools in the studied state were flagged by both Z_0 and the Bayesian procedure at the same time.

Furthermore, the criterion of flagging a school should be higher than the state-level average rate in practice. When the flagging cut increased to 0.2, 99% of schools were flagged by both Z_s and the Bayesian procedure at the same time.

Discussion

Results from the simulation study demonstrated that the proposed easy-to-compute Z_{s} statistic performed well in null and alternative conditions. Specifically, in the conditions tested, Z_{s} had similar probability and precision of detecting aberrant response time patterns as van der Linden & Guo's (2008) Bayesian procedure. On the contrary, the other two simpler statistics proposed in this paper (Z_0 and Z_{l}) do not have adequate power in detecting aberrant response times. Furthermore, it was shown that test lengths, proportion of aberrant examinees in the sample, as well as the number of compromised items all exert an influence on the performance of the indices. van der Linden & Guo (2008) argued that the Bayesian procedure accounts for the presence of estimation error in any of the parameters of the psychometric model (e.g., ability parameters). Our method, on the other hand, eliminated any estimation error because no latent variable needs to be estimated. Moreover, we noticed that the Bayesian procedure had a slightly higher probability of detection than Z_{m}, which is also based on the lognormal response time model. However, the probability and precision of detection of Z_{s} are always close to van der Linden & Guo’s (2008) Bayesian procedure. Another important finding was that, when the test was short (20 items) and the proportion of compromised items was low, all the statistics had a low probability of detecting aberrant response time patterns. The precision of detection was lower or equal to 54%. Therefore, it is recommended that no individual-level detection should be carried out in this situation. Only when the test length is adequate (40 items), are the detection results by the proposed Z_{s} statistic sufficiently reliable. Additionally, empirical data analysis showed that Z_{s} could flag a large proportion of examinees flagged as aberrance by the Bayesian procedure. When the statistics were aggregated at the school level, Z_{s} and the Bayesian procedure almost flagged the same schools even if the aberrance rates were very low in the real data. This finding provided further evidence that Z_{s} would be a useful and powerful statistic in practice.

This study has its limitations. First, the data for the simulation study was generated with a lognormal response time model. This doesn't take into consideration all the features of a real response time data set. For example, test-takers' response time on one item might not follow a lognormal distribution and their response speed might vary during testing. Secondly, only one set of empirical data set was used to illustrate the new statistics. This limits the number of types of aberrant test behaviors we detected in this study. The performance of the proposed statistics need to be examined in more situations. Furthermore, future study should consider improving the current data forensic methods by incorporating more information sources, detecting aberrant examinee behaviors using item response, response time and answer changing patterns.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014*). Standards for educational and psychological testing.* Washington, DC: AERA.

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.

Fox, J. P. & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research, 51(4)*, 540-553.

Fox, J. P. & Marianti, S. (2017). Person-fit statistics for joint models for accuracy and speed. *Journal of Educational Measurement, 54(2),* 243-262.

Impara, J. C., Kingsbury, G., Maynes, D., & Fitzgerald, C. (2005). *Detecting cheating in Computer Adaptive Tests using data forensics.* Paper presented at 2005 Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.

Marianti, S., Fox, J., Avetisyan, M., Veldkamp, B.P. & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, *39(6),* 426-451.

Li, Z., & Smith, J. (2015). *Detecting aberrant behaviors with a hierarchical lognormal response time model.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Chicago, IL.

Qian, H., Staniewska, D., Reckase, M. & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues and Practice*, *35(1),* 38-47.

Reise, S. P. & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement*, *15(3),* 217-226.

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

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*, 181-204.

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics, 33(1),* 5-20.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73(3),* 365-384.

van der Linden, W. J. (2009). A bivariate lognormal response-time model for the detection of collusion between test takers*. Journal of Educational and Behavioral Statistics, 34(3)*, 378-394.

van der Linden, W.J. (2011). Modeling response times with latent variables: principles and applications. *Psychological Test and Assessment Modeling, 53(3)*, 334-358.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, 29(5), 323-339.

Wang, C. & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology, 68*, 456-477.

Wise. S. & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education,* 18, 163-183

Appendix

Table 1

Simulation Study Design

Factors	Conditions
Sample size (N)	500
Number of items (n)	20, 40, 70
Aberrance rates of items (ARI)	5%, 10%
Aberrance rates in sample (ARS)	10%, 25%

*The factors are fully crossed in a 1 × 3 × 2 × 2 design with 500 replications in each condition.

Table 2

Type I error rates and the probability of false alarm

n	Z_0			Z_l			Z_s		
	$\alpha_{-0.025}$	$\alpha_{0.025}$	$R_{flagged}$	$\alpha_{-0.025}$	$\alpha_{0.025}$	$R_{flagged}$	$\alpha_{-0.025}$	$\alpha_{0.025}$	$R_{flagged}$
20	.000	.053	.08	.023	.023	.06	.026	.024	.00
40	.000	.052	.09	.029	.021	.09	.026	.024	.00
70	.000	.052	.11	.027	.022	.10	.026	.024	.00

Table 3

Statistical power at individual level

n	A.R. in Sample	A.R. of Items	Proposed Simple Statistics						Latent Model Statistics			
			Z_0		Z_l		Z_s		Bayes		Z_m	
			P.d.	P	P.d.	P	**P.d.**	**P**	P.d.	P	P.d.	P
20	.05	.10	.08	.06	.08	.06	**.05**	**.49**	.03	.54	.04	.50
		.25	.12	.09	.13	.10	**.52**	**.94**	.46	.97	.54	.95
	.10	.10	.08	.11	.07	.12	**.03**	**.62**	.02	.63	.03	.60
		.25	.11	.17	.12	.18	**.26**	**.95**	.20	.97	.26	.96
40	.05	.10	.13	.06	.11	.07	**.17**	**.75**	.13	.80	.16	.77
		.25	.19	.09	.21	.12	**.86**	**.95**	.86	.97	.88	.96
	.10	.10	.12	.12	.11	.13	**.09**	**.79**	.07	.81	.08	.79
		.25	.17	.17	.20	.22	**.58**	**.98**	.55	.99	.62	.98
70	.05	.10	.11	.06	.13	.06	**.34**	**.90**	.30	.92	.33	.91
		.25	.18	.10	.25	.12	**.97**	**.98**	.97	.99	.97	.98
	.10	.10	.12	.12	.13	.13	**.15**	**.91**	.12	.91	.14	.91
		.25	.16	.17	.23	.20	**.84**	**.99**	.83	.99	.85	.99

Table 4

Percentages of aberrance flagged by different statistics in empirical data analysis

	Aberrance Type	$\boldsymbol{Z_0}$	$\boldsymbol{Z_l}$	$\boldsymbol{Z_s}$	Bayesian procedure	$\boldsymbol{Z_m}$
Flagged response times	All	3.6%	5.0%	5.0%	5.0%	4.9%
	Positive	3.6%	2.8%	2.8%	2.8%	2.8%
	Negative	0%	2.2%	2.2%	2.2%	2.1%
Flagged Examinees	All	3.2%	5.8%	4.5%	4.5%	4.3%

Table 5

Similarity of examinees' flagging with different statistics

	$\boldsymbol{Z_0}$	$\boldsymbol{Z_l}$	$\boldsymbol{Z_s}$	**Bayesian procedure**
$\boldsymbol{Z_l}$	34.8%	--	--	--
$\boldsymbol{Z_s}$	11.8%	45.5%	--	--
Bayesian procedure	11.8%	46.9%	88.4%	--
$\boldsymbol{Z_m}$	11.8%	47.5%	89.6%	93.9%

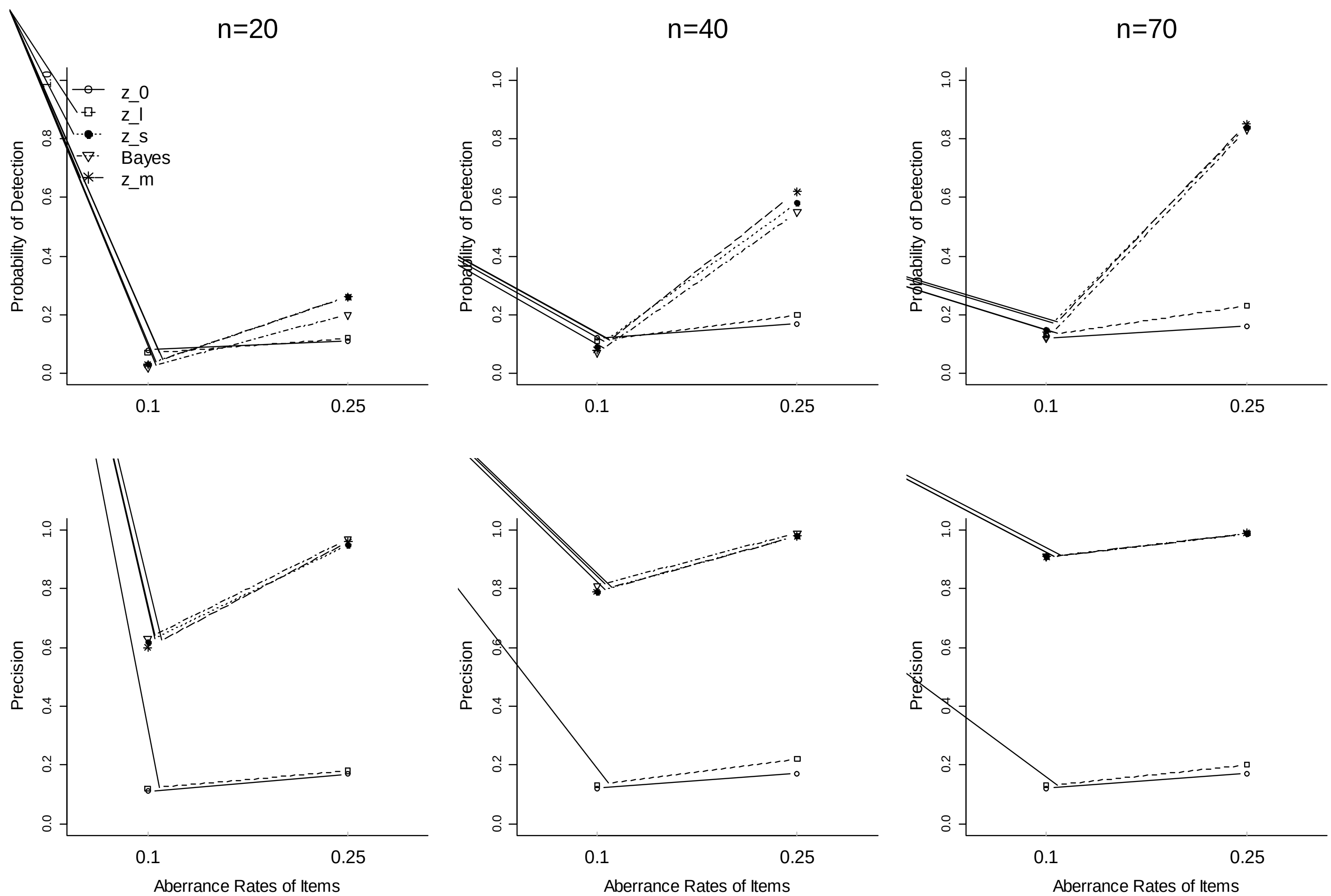
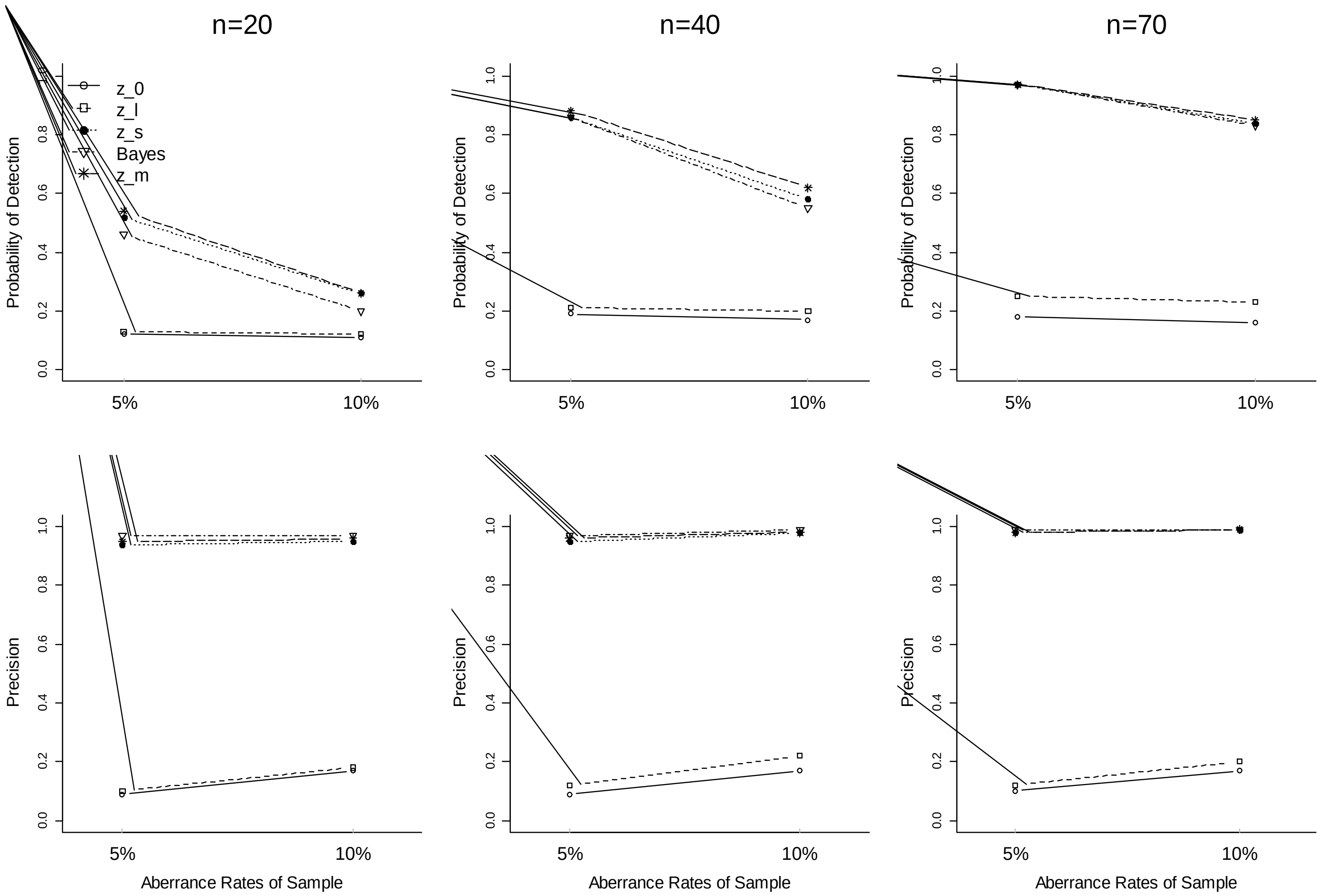


Figure 1. Probability of detection (top) and precision of detection (bottom) of five statistics across test length and aberrance rates of items



1

Figure 2. Probability of detection (top) and precision of detection (bottom) of all statistics across test length and aberrance rates of sample

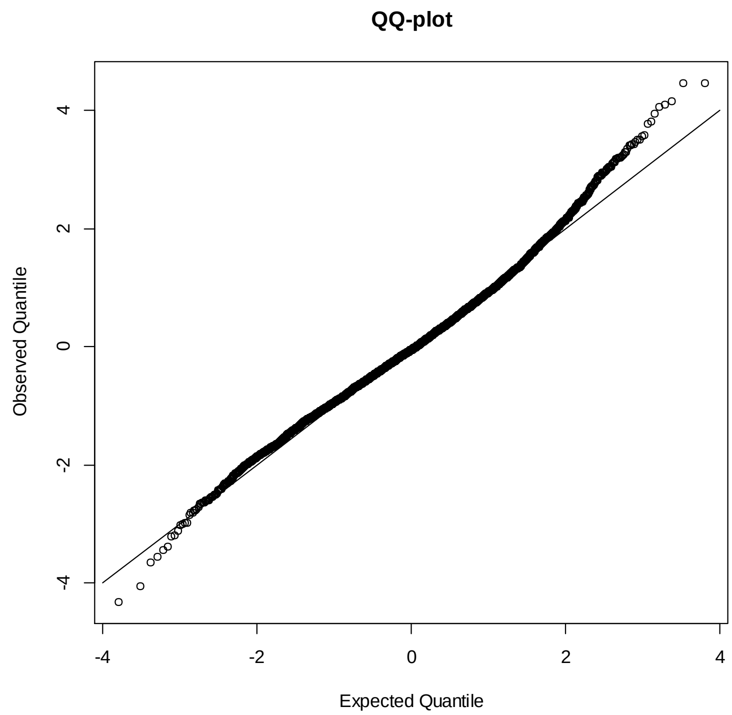


Figure 3. Quantile-quantile plot for total response times on the test in Empirical data analysis

Appendix

Table A1

The likelihoods of randomly flagging a student at or above different cut levels (fixed number of items)

	Number of Items as the Cut Level									
n	1	2	3	4	5	6	7	10	15	20
20	64.15%	26.42%	7.55%	1.59%	0.26%	0.03%	0.00%	0.00%	0.00%	0.00%
30	78.54%	44.65%	18.78%	6.08%	1.56%	0.33%	0.06%	0.00%	0.00%	0.00%
40	87.15%	60.09%	32.33%	13.81%	4.80%	1.39%	0.34%	0.00%	0.00%	0.00%
50	92.31%	72.06%	45.95%	23.96%	10.36%	3.78%	1.18%	0.02%	0.00%	0.00%
60	95.39%	80.84%	58.26%	35.27%	18.03%	7.87%	2.97%	0.07%	0.00%	0.00%
70	97.24%	87.08%	68.63%	46.61%	27.21%	13.72%	6.04%	0.25%	0.00%	0.00%
80	98.35%	91.39%	76.94%	57.16%	37.11%	21.08%	10.53%	0.65%	0.00%	0.00%
90	99.01%	94.33%	83.36%	66.42%	47.03%	29.48%	16.39%	1.45%	0.00%	0.00%
100	99.41%	96.29%	88.17%	74.22%	56.40%	38.40%	23.40%	2.82%	0.01%	0.00%
120	99.79%	98.45%	94.25%	85.56%	72.18%	55.85%	39.37%	7.86%	0.10%	0.00%

Table A1 shows the probability of randomly flagging an individual when the cut increases from 1 to 20 aberrant RTs, for test lengths of 20-120 items. For example, when the test only contains 20 items, an individual with 6 aberrant RTs will have a low false positive rate (0.03%, only 3 students will be randomly flagged in 10,000 students). However, when the test has 70 items, a cut of 6 will have a high false positive rate (13.72%, 1,372 students will be randomly flagged in 10,000 students). By this table, a cut for flagging individuals can be determined after the total number of aberrant RTs for each individual is computed.