Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
xxxx 2016, Vol. 00, No. 0, pp. 1–10

Using Response Time to Detect Item Preknowledge in Computer-Based Licensure Examinations

Hong Qian, *National Council of State Boards of Nursing*, Dorota Staniewska, *Questar Assessment*, Mark Reckase, *Michigan State University*, and Ada Woo, *National Council of State Boards of Nursing*

This article addresses the issue of how to detect item preknowledge using item response time data in two computer-based large-scale licensure examinations. Item preknowledge is indicated by an unexpected short response time and a correct response. Two samples were used for detecting item preknowledge for each examination. The first sample was from the early stage of the operational test and was used for item calibration. The second sample was from the late stage of the operational test, which may feature item preknowledge. The purpose of this research was to explore whether there was evidence of item preknowledge and compromised items in the second sample using the parameters estimated from the first sample. The results showed that for one nonadaptive operational examination, two items (of 111) were potentially exposed, and two candidates (of 1,172) showed some indications of preknowledge on multiple items. For another licensure examination that featured computerized adaptive testing, there was no indication of item preknowledge or compromised items. Implications for detected aberrant examinees and compromised items are discussed in the article.

Keywords: aberrant examinee, item preknowledge, response time, test security

Item preknowledge occurs when some examinees (i.e., test takers) have illegal access to some items in the item pools from previously administered tests prior to taking the examination. Therefore, they can answer these items correctly during the test that they otherwise cannot because they do not have knowledge to answer them correctly. In this article, examinees who have item preknowledge are referred to as aberrant examinees and items being memorized are referred to as compromised items. Item preknowledge can occur in computer-based licensure examinations for at least three reasons. First, substantial stakes are associated with licensure examinations, as the passing of such a test usually makes the examinee eligible to work in a given field, or qualifies the individual for a more highly paid position (Smith & Davis-Becker, 2011). As a result, examinees might be motivated to seek access to some of the items before they take the test. For example, the examinee may ask a classmate who has taken the exam earlier about items they remember, search the Internet for stolen content, or explore the offerings of training schools that provide the content as part of test-preparation materials (Smith & Davis-Becker, 2011). Second, the problem of item exposure is further exacerbated because items in computer-based testing usually remain operational for some time in order to provide a return on the investment in item development (van der Linden & van Krimpen-Stoop, 2003). During this time period, examinees may try to memorize and share items with others. Third, many examinations (including licensure and educational testing programs) are delivered on demand and have large examinee volumes, making them vulnerable to item exposure problems (Cohen & Wollack, 2006).

Item preknowledge can threaten the validity of the inferences from examination scores because it is unclear whether the examinee passed the test because of the knowledge he/she has in the given field or because he/she has item preknowledge. Therefore, it is important for operational licensure testing programs to identify potentially compromised items by monitoring examinee response behavior (Smith & Davis-Becker, 2011).

A traditional way to detect item preknowledge is to conduct person-misfit analysis using response data (McLeod & Lewis, 1999; Meijer & Sijtsma, 1995; van Krimpen-Stoop & Meijer, 2000; Veerkamp, 1996). If a low-ability examinee answered more difficult questions correctly than would be expected by chance, this may indicate that he/she knew these items before he/she took the test. However, this method usually has a low detection rate and a high false alarm rate and cannot be used in operational testing. One possible way to counter these problems is to complement analysis of response data with analyses of response time data.

Hong Qian, Examinations Department, National Council of State Boards of Nursing, 111 E. Wacker Drive, Suite 2900, Chicago, IL 60601-4277; hqian@ncsbn.org. Dorota Staniewska, Questar Assessment, Inc., 5550 Upper 147th Street W, Minneapolis, MN 55124; dstaniewska@gmail.com. Mark Reckase, Michigan State University, East Lansing, MI 48824; reckase@msu.edu. Ada Woo, National Council of State Boards of Nursing, 1111 E. Wacker drive, Suite 2900, Chicago, IL 60601-4277; awoo@ncsbn.org.

© 2016 by the National Council on Measurement in Education

In computer-based testing, each examinee's response time for each item is automatically recorded. Unexpected response times may indicate certain types of aberrant response behavior. For example, examinees who know some of the items prior to taking the test may answer them more quickly than is typically the case. A key assumption in using response time data is that examinees will not fake response time. This is a realistic assumption for two reasons. First of all most examinees do not know that their response time is monitored and used for analysis because most score reports do not report on response time. Second, for students who have item preknowledge, they not only want to answer the current item correctly but also quickly, so they can have more time for other questions since most exams are time-limited. After they identify a familiar item and recall a memorized answer, they would move to the next item and save more time for other questions where they do not have item preknowledge. Therefore, for high-stakes, time-limited testing, it is realistic to assume that the response time to produce an answer to a particular item reflects the time needed to process the item (Meijer & van Krimpen-Stoop, 2010). The purpose of this research was to detect examinees' item preknowledge behavior and detect compromised items in an item pool using response data and response time data for two computer-based operational licensure tests. These two large-volume high-stakes licensure tests use two popular testing administration modes: computer-based nonadaptive test and computerized adaptive testing. Using the real data, we show the application of a theoretically sound procedure in operational testing programs to monitor test security. The significance of this study includes (1) a detailed illustration of how a response-time model could be applied in operational testing to flag potentially compromised items and aberrant examinees in a licensure exam context and (2) an application of the method that can also be used to detect compromised items in educational measurement since some test items are used on more than one occasions.

Specifically, two research questions were addressed in this study:

- To what extent do examinees have item preknowledge in two large-scale operational licensure exams?
- To what extent are there potentially compromised items in each item pool?

Literature Review

Several researchers have tried to use response times to detect possible aberrances in examinee behavior. Van der Linden and van Krimpen-Stoop (2003) used response data and response time data to detect two different types of aberrances: item preknowledge and speededness. They used classical procedures and Bayesian posterior predictive checks in simulation studies. For classical checks, the detection rate was .30 and the false-alarm rate was .05. For Bayesian checks, the detection rate doubled relative to the classical checks, but at the cost of a considerable increase in the false-alarm rate.

Meijer and Sotaridona (2006) used effective response time to detect item preknowledge. Effective response time is defined as the time required for an individual examinee to answer an item correctly. To investigate the power of this method, a random sample of examinees from their data set was selected and their response time was changed to one half or one fourth of the original response time on one half or three fourths of all the items they responded to. This method is sensitive to the amount of time reduced due to item preknowledge. For example, the method has high power to detect item preknowledge for examinees who know half of the items and whose quick responses are equal to one fourth of the normal time (the detection rate is .944). However, it is unrealistic to assume that examinees would have access to half of the items on the test.

Van der Linden and Guo (2008) used a hierarchical framework to detect two aberrant response-time patterns: item preknowledge and taking tests only for the purpose of memorizing the items (unexpectedly long response times). The procedure was illustrated using a data set for the Graduate Management Admission Test (GMAT). The procedure revealed that 1.69% of examinees spent less time than expected and 2.25% of examinees spent more time than expected. These percentages are close to the nominal significance level of the test, which means that the test takers generally behaved quite regularly according to the response-time model and that cheating or item compromise was certainly not a structural problem for the GMAT. Moreover, a power study was conducted using simulated cheating behavior. The response time for known items was set to 10, 20, or 30 seconds. The detection rate for 10 seconds was quite high (.8) and for 20 seconds it was acceptable (.4); but for 30 seconds, the detection rate was low (.2). It should be noted, however, that these rates were only for a single item. If a test taker knows more than one item, the power of the procedure would increase immediately. Since this procedure had a good detection rate, the current research used the procedure from this study (van der Linden & Guo, 2008).

Models Used

This research used a response time model proposed by van der Linden (2006) for modeling response times. For more detailed information, please refer to van der Linden (2006). Suppose test taker j operates on item i at speedτ_j ($\tau_j \in \Re$) and the observed response time is t_{ij},

$$f(t_{ij}; \tau_j, \alpha_i, \beta_i) = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}[\alpha_i(\ln t_{ij} - (\beta_i - \tau_j))]^2\right\}, \quad (1)$$

where $\beta_i$represents the time intensity of item i and $\alpha_i$represents the discriminating power of item i for person speed, similar to the discrimination parameter in item response theory (IRT) but for person ability. The basic idea from this model is that if an item is very time-intensive and you have a low speed, your expected response time on this item is long.

We used a two-parameter logistic (2PL) model for modeling response for a computer-based licensure test which is not adaptive (refers to as Exam 1-Nonadaptive) and used a Rasch model for another computer-based licensure test that features computerized adaptive testing (referred to as Exam 2-CAT).

A software package written in R, CIRT, is available for parameter estimation for both response and response time models (Fox, Entink, & van der Linden, 2007). Parameters are estimated using a Bayesian approach with Gibbs sampler as a Markov Chain Monte Carlo (MCMC) method for sampling from the posterior distribution of the parameters (van der Linden & Guo, 2008). The input files are $N \times K$ matrix of the responses ($N =$ the number of persons; $K =$ the number of items) and $N \times K$ matrix of the log response times. The output files include all model parameters (such as α_i, β_i, and τ_j) and some of the model fit criteria. Then the residuals can be calculated as follows:

$$e_i = \alpha_i(lnt_{ij} - (\beta_{\mathrm{i}} - \tau_j)), \mathrm{e_i}\text{~}N\ (0, 1)\,.$$

Method

Context

To better illustrate the application of the methodology, we used data from two large-scale computer-based licensure exams, one from the financial industry (Exam-1-Nonadaptive) and the other from a health care profession (Exam 2-CAT). These two licensure exams follow the standard item development process established in the licensure testing industry (practice analysis—test plan—item writing—item review—item pretest—exam administration), which makes the results more generalizable to other licensure exams that are in compliance with the standards (AERA, APA, & NCME, 2014). The item type used in these two exams are typical: all multiple-choice items for Exam-1-Nonadaptive and multiple-choice items and alternative item types (including multiple-response, fill-in-blank, and ordered-response) for Exam 2-CAT. Therefore, the methodology detected item preknowledge not only for multiple-choice items but also for alternative item types. Certain percentage of items address a specialized subset of skills obtained from the practice analysis, which is very common for licensure exams.

Each of these two exams is the sole exam that allows the candidates to pursue their chosen career, and therefore a vast majority of exam candidates are known to prepare for the exam by using preparation schools or individual study. A survey of thousands of Exam 2-CAT candidates regarding their exam preparation indicated that about 70% of candidates utilize paid review courses or practice exams either in person or online. Most candidates spent at least 50 hours preparing for the exam outside of an educational program. While it is understandable that owners of preparation businesses would want to garner feedback about the quality of their preparation, it naturally raises concerns about potential item exposure should they ask, for instance, if their client could recall any items verbatim after taking the test. Due to the high-stakes nature of licensure exams, several efforts have been made to minimize the effects of item compromise such as item exposure control, creating a large item pool, and replacing the item pool frequently. Some testing organizations even hire web patrollers to monitor potential threats regularly. For example, Exam 2-CAT has two contractors to handle domestic website searches and international web patrols, respectively, and a weekly report sent back by each contractor is reviewed by exam security team. The methodology illustrated in this article provides an additional way to monitor the item security in addition to all other security efforts that are already in place for most licensure testing organizations as well as to make sure whether those security mechanisms work or not. In the constant battle between those who willingly commit test fraud and the valuable assets a testing program spends time and money to build, an additional protective barrier to ensure the validity of the assessments is of value (ITC, 2014).

Sample

To detect item preknowledge, two samples are needed. The first sample, including an $N \times K$ matrix of the responses and an $N \times K$ matrix of the log response times, is used for item calibration. This sample should be from the early stage of the operational test, where there are no compromised items. The second sample is from the later stage of the operational test, which is subject to a content exposure problem. Data from this sample are used to estimate person parameters since all item parameters are assumed to have been known from previously calibrated items.

For Exam 1-Nonadaptive, two nonconsecutive years of data were used—the first and third years of exam administration. For the early sample, 992 candidates taking a 185-multiple-choice-item exam in the first 6 months of test administration (January to June of 2010) were used to estimate the model. It was assumed that, as those candidates had taken a completely new exam, there was no compromised content and the estimated parameters were the true item parameters. These parameters were then applied to the second sample of 1,172 candidates taking the test in early 2012 to detect possible item preknowledge. There were 111 items in common between the two samples. Slightly over 10,000 candidates took the test between June 2010 and December 2011.

For Exam 2-CAT, each item pool remained operational for 3 months. There were 51,480 examinees who took the test for the first time throughout 3 months from April 2012 through June 2012. The item pool included 1,472 operational items. We only examined the items with at least 100 responses to ensure estimation accuracy, which resulted in inclusion of 1,055 items. The study used the first-month examinees (i.e., April examinees) as the early example and the last-month examinees (i.e., June examinees) as the late example. There were 4,675 first-time examinees in April and 4,604 first-time examinees in June. This is a variable length CAT exam. The minimum test length is 60 items while the maximum test length is 250 items. The average test length is around 110 items. There are three stopping rules for this exam: (1) 95% Confidence Interval Rule—this is the most common stopping rule for examinees. The computer will stop giving items when a 95% confidence interval of ability estimation is clearly above or below the passing standard; (2) Maximum-Length Exam Rule—when an examinee's ability is very close to the passing standard, the computer continues to administer items until the maximum number of items is reached, which is 250 items; and (3) Ran-Out-Of-Time Rule—test takers have 6 hours to complete this exam.

Procedures

Unexpected short response time is detected by estimated residual log response time:

$$a_i(Int_{ij} - (\beta_i - \tau_j)), i = 1, ..., K; j = 1, ..., N.$$

These residuals have an approximate standard normal distribution. Because the speed parameters are estimated for the full test, an increase in the actual speed on subsets of items manifests itself by larger negative values for the residuals. A response time to an item was flagged as aberrant when

© 2016 by the National Council on Measurement in Education

its residual had a larger negative value than –1.96. It is important to acknowledge that there are many related significant tests and one could use a Bonferroni correction (using a much smaller critical value, which would indicate a smaller number of flagged response times). We used –1.96 to be more conservative. Then, we summarize how often the response times for the same person were flagged to detect suspect examinees. We also summarize how often the response times on the same items were flagged to detect compromised items.

Model Validation

There are several assumptions underlying the models used in this study. Before using the models to predict reasonable response times, compare them to the observed ones, and identify unexpected ones, it is necessary first to test whether the data fit the model. The main assumptions underlying the models described in the previous section include the following:

1. The responses fit the IRT models

We used IRTPRO (Cai, Thissen, & du Toit, 2011) to check whether the response data from Exam 1-Nonadaptive fit the 2PL model according to item-level diagnostic statistics. The results indicated that only two items among 111 items showed misfit, which indicates that most of the responses fit the 2PL model.

For Exam 2-CAT, all items included in the operational item pool fit the Rasch model. In other words, if a pretest item did not fit the Rasch model, it was deleted and not included in the operation item pool. Therefore, this assumption was already fulfilled.

1. The response times fit the lognormal distribution

Figure 1 shows a histogram of the response times for one random item from Exam 1-Nonadaptive. The x-axis presents the response times in seconds while the y-axis represents the number of examinees. As is typical of response-time distribution, the data are unimodal and positively skewed. In order to decide if response times for each item come from a population with a lognormal distribution, we used the Kolmogorov–Smirnov goodness-of-fit Test. The results showed that the response times for most items (106 of 111 for Exam 1-Nonadaptive and 1,047 out of 1,055 for Exam 2-CAT) fit the lognormal distribution.

Results for Exam 1-Nonadaptive

Descriptive Data Analysis

Figure 2 shows the distributions of response times in seconds for one item from the 2010 and 2012 administrations of Exam 1-Nonadaptive. The darker bar indicates the right answer while the lighter bar indicates the wrong answer. While the general pattern follows the typical response-time distribution, there is a little bump at the left for each distribution. These are the examinees who responded to the item very quickly. For 2010, about 40 examinees responded to this item in less than 40 seconds while the median time was 400 seconds. It is possible that some examinees did not want to spend a lot of time on one item and quickly guessed an answer and went to the next item. Almost half of the examinees got the item wrong. In 2012, there were even more examinees responding to this item quickly. However, a close look reveals that the increase is greatest in the white part of Figure 2, which refers to wrong answers. Therefore, even though there were more examinees answering the item quickly in 2012 than in 2010 this was not due to item preknowledge, because if an examinee knew an item before the test he/she would have answered it quickly and correctly. While Figure 2 does not reflect item preknowledge, Figure 3 may provide another story.

In Figure 3, the number of examinees in the first bar increased in 2012 and most of them answered the item correctly. This may indicate that some examinees knew this item before they took the test, given that the item had been exposed for 2 years. Figure 3 provides a visual representation of a possibly compromised item. The statistical procedure used in this study will identify such items more efficiently and accurately than such plotting.

The procedure used in this study can not only detect compromised items, but also identify examinees who answer some questions unexpectedly quickly relative to his/her own speed. For an examinee who is regularly fast, the response time needs to be extremely short in order to be detected as unexpectedly quick while for an examinee who is extremely slow, a typical response time may seem quick for this examinee. Figure 4 displays the fastest examinee's response times for 111 items against the median response times for Exam 1-Nonadaptive.

For the fastest examinee in Figure 4, his/her response times for almost all items are shorter than the median response times. At the same time, this examinee also follows the pattern of response times for each item, which means that he/she spent a little bit more time on the items that others generally spent a lot of time on. Figures 5 presents the residuals for the same examinee.

In Figure 5, the x-axis features 111 items, and the y-axis represents the residuals. A residual above 0 means the examinee spent more time on this item than expected based on the person's speed and the time intensity of the item. A residual below 0 indicates the examinee spent less time on this item than predicted from the model. The solid line indicates the right answer while the dash line indicates the wrong answer. For the fastest examinee, it is not surprising that there are a lot of negative residuals. However, among those negative residuals, none are less than –1.96. Figure 5 shows that this examinee was regularly fast and not extremely fast on some items. Therefore, there is no indication of item preknowledge.

Results for Detecting Item Preknowledge and Compromised Items

We checked the residual log-response time for all examinees in the 2010 and 2012 samples. The reason for checking residual log-response time for examinees in the 2010 sample is that the number of flagged response times in 2010 reflects a type 1 error because in 2010 there should have been no item preknowledge cases. Then the study compared the number of flagged response times for examinees in the 2010 and 2012 samples. Using flagged response times in 2010 as a baseline, we were able to control for the type 1 error. The results showed two examinees in 2012 who had significantly higher numbers of flagged items than the 2010 baseline. Figure 6 displays the residuals for one of the examinees.

Note that the large negative residuals indicate faster responses than typical for this examinee. The examinee in Figure 6 spent an extremely short amount of time on five items (items 3, 9, 24, 52, and 80) relative to the time he/she spent on the other items. For example, the residual log-response time for item 24 was –4.1. As the model assumed a standard

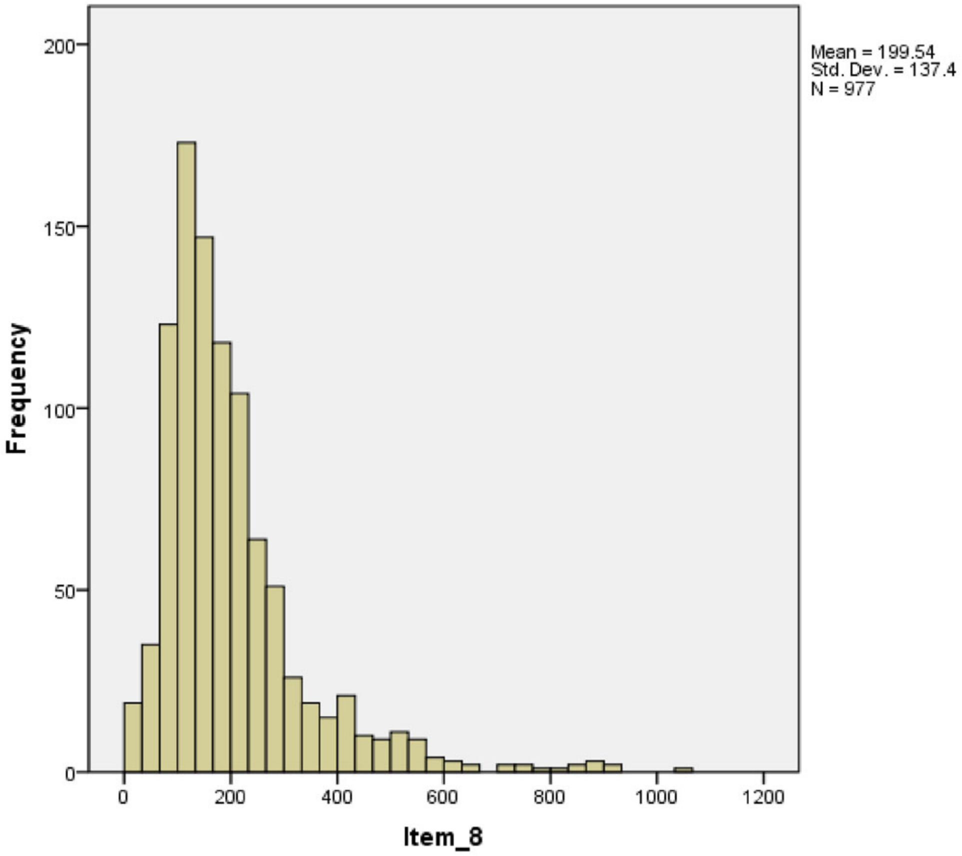

FIGURE 1. A typical histogram of response times from one random item.

normal distribution of the residuals, the probability of such a negative residual by chance alone is .000021. The estimated time intensity of this item was $\beta = 4.63$, which means that the median time for a test taker of average speed ($\tau = 0$) taken on the item was 102 seconds on the regular time scale. This examinee spent 3 seconds on this item even though the individual's estimated speed is lower than average overall ($\tau = -.14$). The examinee responded to all five of these items correctly and only responded correctly to 60% of the remaining 106 items. As these items are spread throughout the test, the examinee's pattern is consistent with possible preknowledge of these five items rather than rapid guessing due to speededness or loss of motivation. Another examinee spent an extremely short time on three items and all of them were correct, which may indicate the preknowledge of these three items.

To evaluate whether there were compromised items in the 2012 sample, the study compared the number of flagged response times for each item for the 2010 and 2012 samples using a paired t-test. The result showed that there were significantly more flagged response times for the 2012 items than for the 2010 items ($t = 2.943, p = .004$).

After looking at every item in 2010 and 2012, two items showed a significant increase in the number of flagged residuals. Figure 7 shows one of the two potentially compromised items that the study discovered. For this item, there were 34 flagged response times, 26 of which (76.5%) were correct. According to Wise and Kong (2005), the accuracy of the rapid-guessing responses should not exceed the level of chance, as this clearly does. The results indicate that the item may have been compromised. An additional examination of the item's content reveals a long stem with a memorable story, which might add to the ease of knowledge transfer between examinees.

In conclusion, after comparing with the baseline in the first year, we found item preknowledge in the third year to be minimal, with two items (out of 111) potentially exposed, and two candidates (out of 1,172) showing some indication of preknowledge on multiple items for Exam 1-Nonadaptive. To make sure that the results were not due to a lack of power of the method, a simulation study was conducted using short response times from one of the detected compromised items to determine whether the method can accurately identify these cases. The result showed that, when examinees know 10% of the items or less, the procedure can detect 67 out of 100 "true" item preknowledge cases (Qian & Staniewska, 2013). Furthermore, a cross-validation of the results showed that there is significant item difficulty drift from 2010 to 2012 for these two compromised items detected by the response time model. Another confirming finding is that two aberrant examinees responded to these two compromised items quickly and correctly.

Results for Exam 2-CAT

For the April sample, among 461,949 item–person combinations, 139 item–person combinations' residuals were smaller than –1.96. For the June sample, among 420,268 item–person combinations, 35 item–person combinations' residuals were

© 2016 by the National Council on Measurement in Education

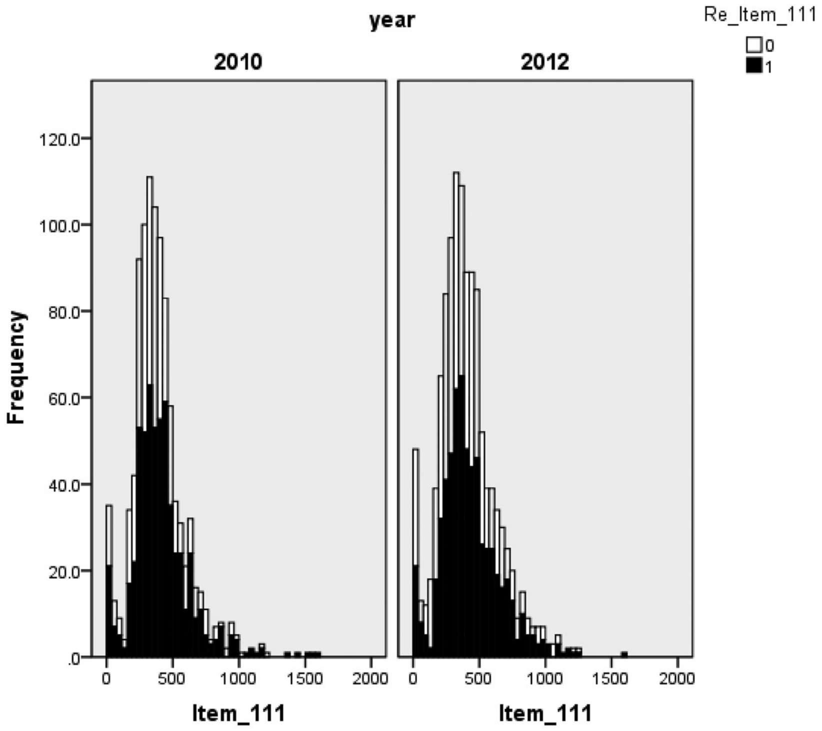

FIGURE 2. The distributions of response times in seconds for item 111 in 2010 and 2012.

smaller than –1.96. The reason for less flags in the June sample is that most of the examinees had just graduated and had higher ability and speed than the April sample. According to the analysis results, there is no evidence of item preknowledge for the 3-month period of item exposure for Exam2-CAT. This result may be due to several characteristics of this large-scale, computerized adaptive licensure test. First, this test has a large item pool. The master pool contains more than 14,000 items while the operational item pool for each 3-month period includes 1,472 items. It is difficult for examinees to memorize such a large number of items. Second, this exam uses randomesque with 15 items as exposure control, which means that even if two examinees had the same ability, they only had a 1/15-chance of seeing the same item. Therefore, the item overlapping rate is too low to make sharing the item useful. Third, potential online threats are monitored regularly by two contractor companies and are reviewed by the exam security team. Based on these characteristics, the result from this study is not surprising.

Among 139 flagged residuals for the April sample, the number of flagged residuals for each item and the number of flagged residuals for each examinee were checked.

One hundred and thirty-nine flagged residuals were distributed on 119 items. For 115 items, there were 1 or 2 flags for each item. For four items, there were 3 flagged residuals for each item. It seems that these items were not compromised. If 139 flagged residuals were all on one item, then this item may have been compromised and should be deleted from the item pool. But this is not the case for this study.

One hundred and thirty-nine flagged residuals were distributed on 31 examinees. For 27 examinees, there were a few flags for each examinee. However, for the remaining four examinees, there were 31, 23, 22, and 21 flagged residuals, respectively. Figure 8 displays the residuals for one examinee with 31 flagged residuals.

As shown in Figure 8, this examinee spent more time than expected on the first 120 items. Then he/she spent a little less time than expected for 200 items. Finally he/she changed the pattern and responded very quickly on the rest of the items. A lot of residuals were smaller than –1.96. However, among them most responses were incorrect (dash lines). The proportion of correct responses was about what was expected by chance. Therefore, even though there were a lot of large negative residuals, they were not an indication of item preknowledge since most of them were wrong. It is possible that the examinee lost motivation because the test was too long (he/she did not run out of time, so it was not a case of speededness). If this examinee lost motivation because the test seemed endless, then his/her final ability estimation is not accurate due to rapid guessing toward the end of the test. It is reasonable from a psychometric perspective to conclude that if the test taker's ability is near the cut score, the computer should continue to administer items. However, this may influence examinees' psychological status and lead to aberrant response patterns. If there are a lot of examinees showing such patterns, the stopping rule may need to be modified. Among 4,675 examinees, 4 examinees showed a similar pattern. Therefore, the effect was minimal for this test.

Conclusions and Implications

Checking the response time of test takers for possible aberrant behaviors is made possible by administering the test on the computer (van der Linden & Guo, 2008). As argued in this article, item preknowledge can be a key component of the information related to the validity of the inferences from test results. We used a response-time model proposed by van der Linden (2006) to address two research questions: (1) To what extent do examinees have item preknowledge in two

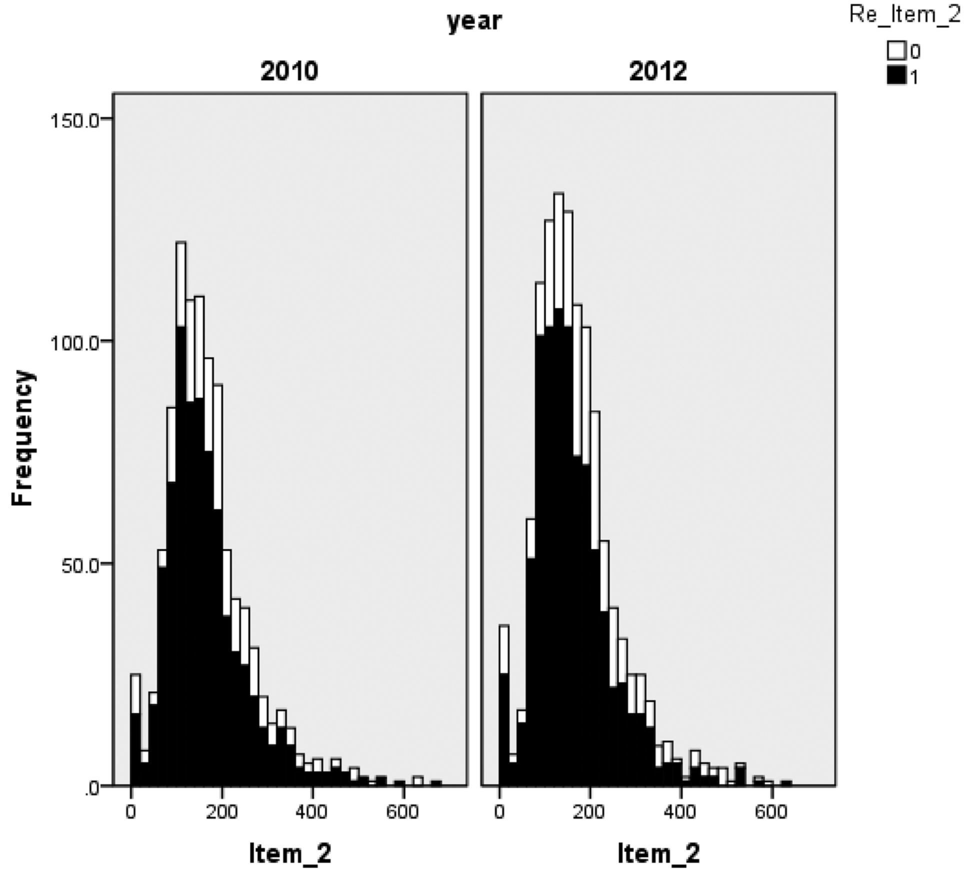

FIGURE 3. The distributions of response times in seconds for item 2 in 2010 and 2012.

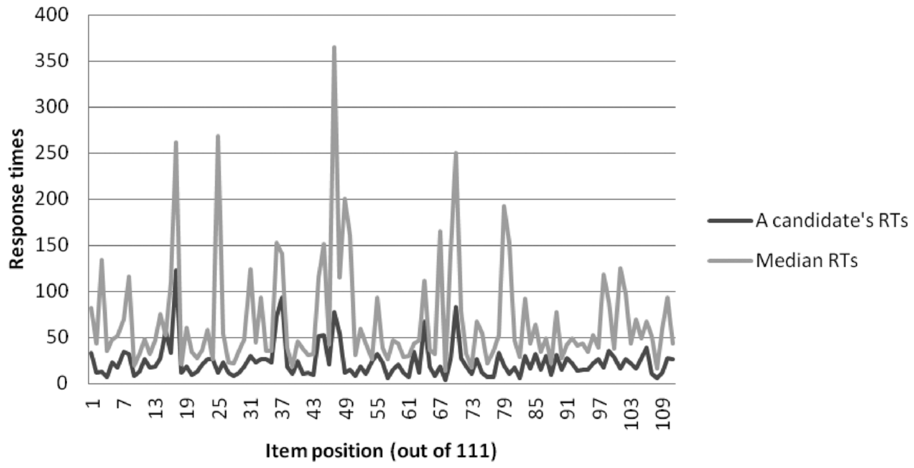

FIGURE 4. The fastest person's response times against median response times from Exam 1-Nonadaptive.

large-scale operational licensure exams? (2)To what extent are there potentially compromised items in each item pool? For Exam 1-Nonadaptive, the results showed that two items (out of 111) were potentially exposed, and two candidates (out of 1,172) showed some indication of preknowledge on multiple items. For Exam 2-Adaptive, there was no indication of item preknowledge or compromised items. However, a limitation of the model is that it could not detect item preknowledge if aberrant examinees faked realistic response times. Examinees who know that the response times are monitored might alter their test-taking behavior by slowing their response, such as reading the stems and options multiple times. Therefore organizations need to evaluate whether faking response time is possible before using the model to monitor testing security.

Our results have different implications for detecting (1) examinees with possible item preknowledge and (2) compromised items. For examinees with possible item preknowledge, the results need to be used with caution. Even though the simulation study (van der Linden & Guo, 2008;

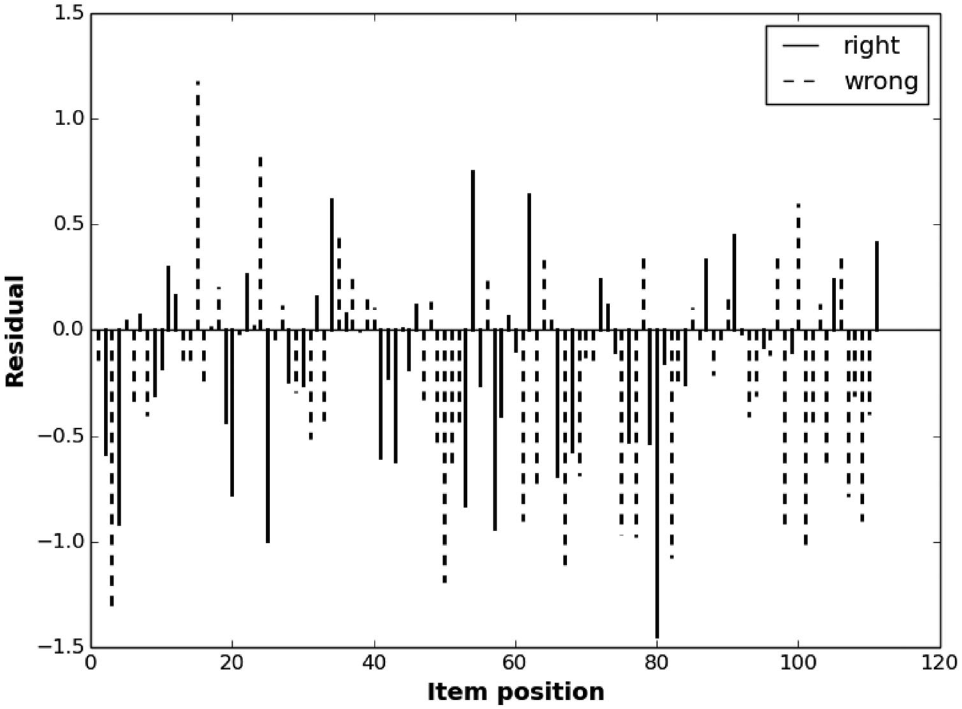

FIGURE 5. The fastest examinee's residuals from Exam 1-Nonadaptive.

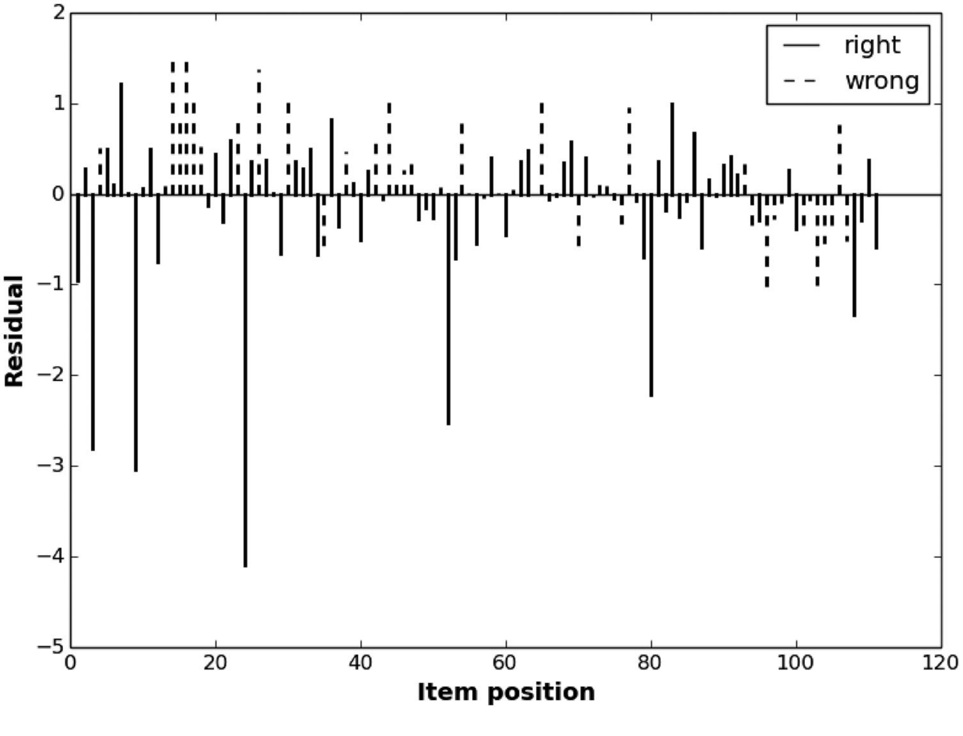

FIGURE 6. Residuals of an examinee with possible item preknowledge from Exam 1-Nonadaptive.

Qian & Staniewska, 2013) showed satisfactory power for the proposed checks, there are other explanations for aberrant response times besides item preknowledge, and blind conclusions from statistically significant log response time residuals could easily be wrong (van der Linden & Guo, 2008). After detecting examinees with possible item preknowledge, careful qualitative analyses are needed, such as a review of the reported irregularities during testing sessions or an investigation of video recordings of test takers while they were taking the test. The analysis of testing center behavior could be helpful if item preknowledge is suspected because this type of finding typically indicates that a test taker had information memorized (as they moved through the item quickly) as compared to cheating within the room. The

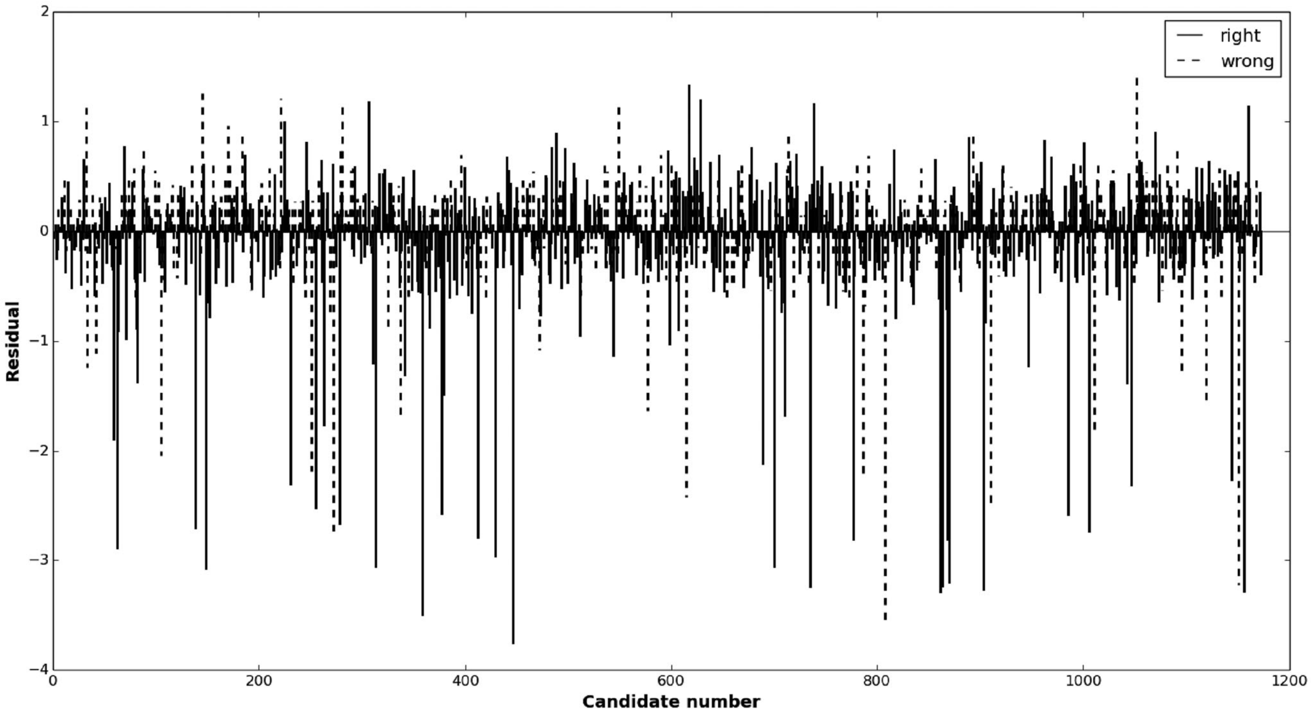

FIGURE 7. Residuals of a possibly compromised item from Exam 1-Nonadaptive.

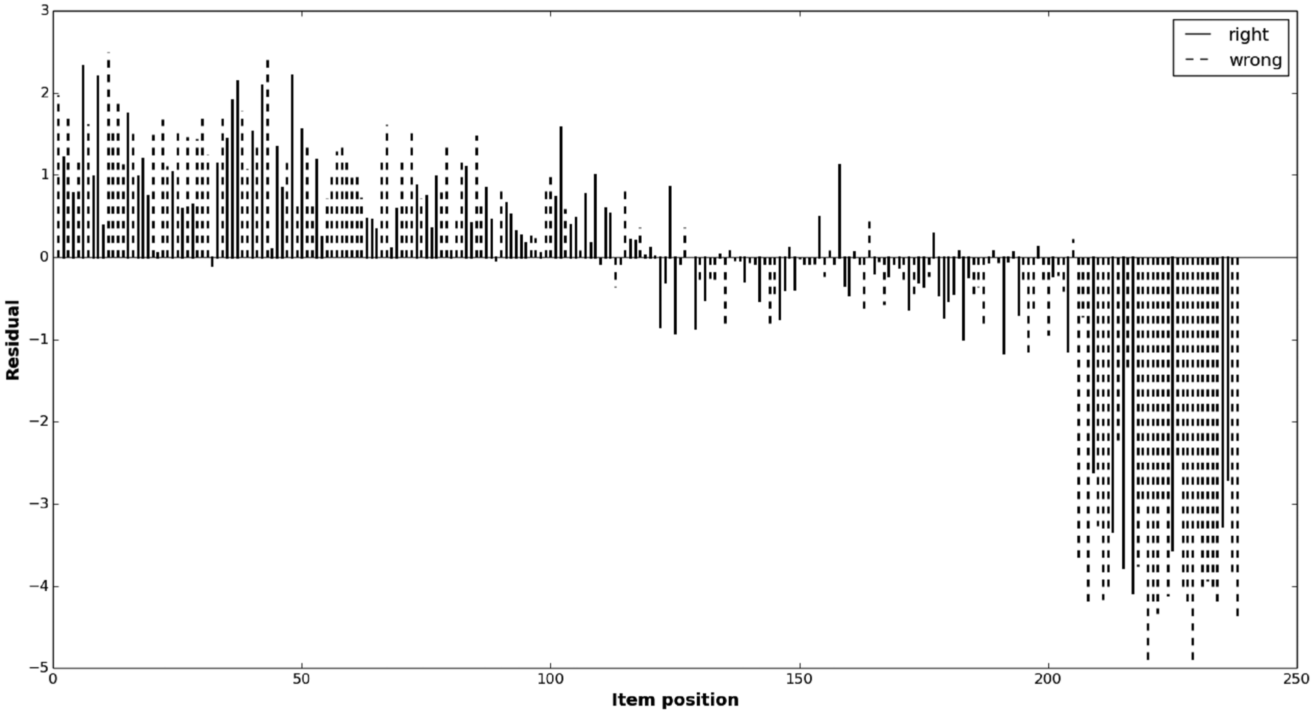

FIGURE 8. Residuals for an examinee with 31 flagged residuals (theta = –.37) from Exam 2-CAT.

evidence from psychometric analysis alone is not strong enough to invalidate the test takers' test scores, but it is a good starting point to trigger further investigation. For testing organizations with large testing volume every day, this psychometric screen can help reduce human workload to a large degree.

For compromised items that are detected, a more conservative attitude can be taken. The simplest response is to delete these items from the item pool and never use them in an operational test again. It is possible that there are some false alarms, but it is wise to delete an item rather than taking a risk. Furthermore, the implications from this study can help prevent item preknowledge at the item development stage, too. For example, after detecting the compromised items, it is necessary to examine the content of these items and identify characteristics of them that are more easily exposed to the test takers' population.

For the two items detected in this study, the actual content was examined by content experts and it was determined that these two items are very memorable. For one item, there was a long story in the stem that is easy to memorize and communicate. For another item, there was an uncommon phrase put in quotation marks. Therefore, item development guidelines can be devised based on these characteristics to prevent possible item preknowledge in item development. If a long story or an uncommon phrase is not necessary for measuring the construct, it would be better not to include it in the item stem or options to minimize memorability and communication.

This research provided two examples for the application of a theoretically sound procedure in operational testing programs to monitor test security. Our results can not only help these two testing programs to improve (or ensure) test security by deleting compromised items and further investigating aberrant examinees; it can also help them to regularly check for item preknowledge and enhance item writing in the future. This research also provides an option for other licensure testing programs to monitor item preknowledge or other aberrant examinee behaviors such as speededness or loss of motivation (as indicated from Exam 2-CAT) using response time data. Furthermore, the method can be used in educational measurement practice since some items are used repeatedly and it is possible to use item response time to detect whether some items are compromised.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Cai, L., Thissen, D., & du Toit, C. S. H. (2011). *IRTPRO for Windows* [Computer software]. Lincolnwood, IL: Scientific Software International.

Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring and reporting. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 355–386). New York, NY: Macmillan.

Fox, J.-P., Entink, K. H. R., & van der Linden, J. W. (2007). Modeling of responses and response times with the package CIRT. *Journal of Statistical Software*, *20*(7), 1–14.

International Test Commission (ITC). (2014). *ITC guidelines on the security of tests, examinations, and other assessments*. Retrieved January 27, 2015, from www.intestcom.org.

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

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education*, *8*(3), 261–272.

Meijer, R. R., & Sotaridona, L. S. (2006). *Detection of advance item knowledge using response times in computer adaptive testing*. Law School Admission Council Computerized Testing Report, 03-03. Newtown, PA: Law School Admission Council.

Meijer, R. R., & van Krimpen-Stoop, E. M. L. A. (2010). Detecting person misfit in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp.289-314). New York, NY: Springer.

Qian, H., & Staniewska, D. (2013, April). *Using response time to detect item pre-knowledge in computer-based testing*. Paper presented at the meeting of National Council on Measurement in Education, San Francisco, CA.

Smith, R. W., & Davis-Becker, S. L. (2011, April). *Detecting suspect examinees: An application of differential person functioning analysis.* Paper presented at the annual conference of the National Council on Measurement in Education, New Orleans, LA.

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., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*, 365–384.

van der Linden, W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant response patterns in computerized adaptive testing. *Psychometrika*, *68*, 251–265.

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 & C. A. W. Glas (Eds.), *New developments in computerized adaptive testing: Theory and practice* (pp. 201–219). Boston, MA: Kluwer-Nijhoff Publishing.

Veerkamp, W. J. J. (1996). *Statistical methods for computerized adaptive testing* (Unpublished doctoral dissertation). University of Twente, The Netherlands.

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