Two Statistical Tests for the Detection of Item Compromise

Wim J. van der Linden
University of Twente, Enschede, the Netherlands

Two independent statistical tests of item compromise are presented, one based on the test takers’ responses and the other on their response times (RTs) on the same items. The tests can be used to monitor an item in real time during online continuous testing but are also applicable as part of post hoc forensic analysis. The two test statistics are simple intuitive quantities as the sum of the responses and RTs observed for the test takers on the item. Common features of the tests are ease of interpretation and computational simplicity. Both tests are uniformly most powerful under the assumption of known ability and speed parameters for the test takers. Examples of power functions for items with realistic parameter values suggest maximum power for 20–30 test takers with item preknowledge for the response-based test and 10–20 test takers for the RT-based test.

Keywords: continuous testing; fixed-form testing; item compromise; item response theory; lognormal response-time model; statistical hypothesis testing

Introduction

Attempts to cheat on tests have always existed, specifically for tests used for admission to educational programs or the licensing of candidates for professional practice. Unlike such tests as for individual counseling or diagnostic purposes, where accurate scoring typically is in the interest of all concerned, higher scores typically result in better educational or vocational opportunities. Some of the test takers may therefore feel tempted to try profiting from cheating.

The replacement of group-based, paper-and-pencil testing with online continuous testing with immediate scoring of the test takers has effectively ended traditional forms of cheating as attempts to copy answers from fellow test takers, collude with them through secret forms of communication, or bribing proctors to improve the answer sheet. However, precisely because of its continuous format, it has also led to a new type of cheating in the form of attempts to harvest test items and share them with future test takers, a possibility supported by the recent technological trend toward miniaturization of electronic recording devices and their integration in wearables. In fact, the practice has become so profitable that
organized crime has entered the testing business operating through networks of henchmen and selling items through websites.

Testing programs have been keenly aware of this danger of item compromise and introduced a variety of counter measures, such as monitoring websites and social media for the presence of stolen items, practicing zero tolerance with heavy penalties for test takers caught stealing them, controlling the exposure rates of their items by rotating the item pools, or introducing randomized item-selection technique, as well as the use of forensic statistical analysis.

The idea of forensic analysis already has some tradition. As a matter of fact, a review of the existing literature shows quite a few different statistical methods introduced to detect items likely to be compromised. The first distinction it reveals is between methods designed to identify items that have been stolen and test takers who profited from knowing them. Although related, the two approaches differ in that the focus of the former is on patterns of information about items for a given test taker, whereas, conversely, the latter are primarily based on patterns across test takers for a given item. Another distinction is between the type of information that is used. Earlier methods were exclusively based on the responses produced by the test takers, but more recently, it has become clear that their response times (RTs) are an important additional source of information. Another relevant distinction is between the statistical nature of the methods, which has ranged from methods generally classified as belonging to the areas of statistical quality control, model fit analysis, change-point analysis, and residual analysis to statistical hypothesis testing. Also, several of these methods have been developed along the lines of frequentist statistics, while others follow a Bayesian approach. Finally, the methods may differ in the type of prior knowledge about the testing process they require. For instance, methods to check whether test takers had preknowledge of some of their items typically assume the set of items that were stolen from the pool to be already known.

One of the first proposals to detect items being compromised rather than test takers having profited from them was based on the cumulative sum technique from statistical quality control (Veerkamp & Glas, 2000). The method periodically re-estimates the difficulty parameter of an item and checks the cumulative changes in the estimates until it becomes too easy given the standard error of estimation. The version of the same method by van Krimpen-Stoop and Meijer (2001) can be used to detect test takers with preknowledge of an item. A related sequential procedure is change-point analysis. The first to use it was Zhang (2014) who used the marginal probability of success on the item under an item response theory (IRT) model as a statistic to detect the point of time, at which it begins to increase for an assumed population of test takers during sequential testing. Later versions of the same type of analysis have been proposed by Sinharay (2016, 2017a) and Zhang and Li (2016).

McLeod et al. (2003) were the first to use a Bayesian approach. Their method used a posterior log-odds ratio based on the response probabilities for a test taker.
being in the alternative hypothetical states of already knowing or not knowing the item. A later Bayesian method in the same vein was presented by Belov (2016) and Belov and Armstrong (2011), who used the posterior shift in the test taker’s ability parameter given the two subsets of secure and compromised items (for the prerequisite detection of compromised items, see Belov, 2014). X. Wang et al. (2017) also adopted the idea of comparing the test takers’ performances on suspected and secure items but used predictive distributions estimated from secure items to evaluate the observed responses on the suspected items.

The idea to use RTs rather than the responses on the test items was introduced in van der Linden and Guo (2008). Their method focused on plots of Bayesian residuals across the items for each of the test takers left after posterior prediction of their RTs; for an extensive empirical illustration of the method, both for adaptive and fixed-format testing, see Qian et al. (2016). Alternatively, a posterior expected likelihood-based person-fit statistic proposed by Marianti et al. (2014) can be used to identify test takers with aberrant RT patterns pointing at item preknowledge. Using a similar residual approach, but then from a frequentist perspective, X. Wang and Liu (2020) presented a standardized statistic with known asymptotic mean and variance to identify both items that are compromised and test takers with preknowledge of them.

An entirely different perspective has been offered by Segall (2002, 2004). His approach was not so much aimed at the detection of compromised items or test takers profiting from them but at estimating the gains in score distributions due to the presence of such items in an item pool for adaptive testing or, more effectively, minimizing such gains by adjusting the item-selection rules.

Sinharay (2017b) used the Neyman–Pearson framework of hypothesis testing to introduce a likelihood ratio and score test of item preknowledge. The tests assume the compromised items to be known. The hypotheses tested against each other were whether or not the test taker’s ability parameter in the three-parameter logistic (3PL) response model estimated from the subset of compromised items was higher than for the secure items. Recently, Sinharay and Johnson (2020) extended the likelihood-ratio test to a test of the joint hypotheses of both the test taker’s ability parameter in the response model and speed parameters in the lognormal RT model being equal or different for the two different subsets of items.

This article presents two independent statistical tests of item compromise, one based on the test takers’ responses and the other on their RTs. The tests thus share their use of both sources of information with Sinharay and Johnson but differ in several other aspects. For instance, they are tests of item compromise, not whether or not individual test takers have profited from knowing such items. Consequently, their focus is on the responses and RTs of the test takers for a given item rather than across items for a given test taker. The null and alternative hypotheses for the two new tests are formulated with the number of test takers with preknowledge of the item as unknown parameter. They are tested against
each other using simple, intuitive statistics as the number of correct responses and the total time by the test takers on the item. As will be shown later, though computationally simple, the two tests demonstrate high power for items with parameter values typically met in the practice of educational testing. In fact, they can be shown to be uniformly most powerful (UMP) under the assumption of known ability and speed parameter. Due to these features, a natural application is in real-time monitoring of test items for possible compromise in a continuous online testing program. The application is possible for programs regardless of their testing format (adaptive, fixed format, linear on the fly, etc.); the only thing necessary is the collection of the responses and RTs for a window of test takers when the program is active. The choice of whether the window should be periodic or moving is to be based on practical considerations. It is also possible to use the tests as part of post hoc forensic analysis in a group-based, fixed-format testing program, but the application then misses the agility and immediate flagging of item compromise they offer in a continuous testing environment.

This article is organized as follows: First, the examples of the response and RT models used in the testing program are presented and the basic null and alternative hypotheses of an item being compromised are introduced. The hypotheses are then specified in terms of the model parameters and the two statistical tests are derived, first the test based on the test takers’ responses and then the one based on their RTs. Next, the examples of the power functions for the two tests for the case of known ability and speed parameters are presented for a realistic choice of values for the item parameters. We then discuss the impact of the necessity to estimate the ability and speed parameters in real-world applications. This article concludes with a brief discussion of a few remaining topics.

Models

To introduce the models for the response-based and RT-based tests, the following notation is used. During testing, each item is supposed to be checked for possible compromise using a window of $p = 1, ..., P$ test takers. Let $i$ denote an arbitrary item that is checked. Each of the $P$ test takers produces both a response and an RT on the item, which are represented by random variables $U_{pi}$ and $T_{pi}$, respectively. In addition, we use

$$X_{pi} \equiv \sum_{p=1}^{P} U_{pi}$$

and

$$T_{pi} \equiv \sum_{p=1}^{P} T_{pi}$$
van der Linden
to denote the total number of correct responses by the \( P \) test takers and the total
time they spent on the items, respectively.

For dichotomous items, the distribution of \( U_{pi} \) is Bernoulli with probability
mass function (pmf)

\[
f(u_{pi}; \pi_{pi}) = \pi_{pi}^{u_{pi}} (1 - \pi_{pi})^{1-u_{pi}},
\]

where \( \pi_{pi} \) is the probability of a correct response for test taker \( p \) on item \( i \). The
probabilities are assumed to follow the well-known 3PL model

\[
\pi_{pi} \equiv c_i + (1 - c_i)[1 + \exp(-a_i(\theta_p - b_i))]^{-1},
\]

with \( \theta_p \) the parameter for the ability of the test taker, \( b_i \in \mathbb{R} \) and \( a_i \in \mathbb{R}^+ \) parameters for the difficulty and discriminating power of the item, and \( c_i \in (0, 1) \)
representing the height of a lower asymptote to the response probability used to
account for the effect of guessing. The adoption of the 3PL model is for pre-
sentation purposes only. Any model with separate item and test taker parameters
that explains the probabilities \( \pi_{pi} \) in (1) can be used. The item parameters are
assumed to have been estimated during earlier item calibration with enough
precision to treat them as known, an assumption with generally minor con-
sequences (e.g., Cheng & Yuan, 2010; Liu & Yang, 2018; Yang et al., 2012). The
statistical test will first be presented for the case of known ability parameters,
after which the impact of their estimation on the power of the test will be
evaluated.

The model for the RTs is the lognormal model, which postulates the distri-
bution of the RTs \( T_{pi} \) to have probability density function (pdf)

\[
f(t_{pi}) = \frac{\gamma_i}{t_{pi} \sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left[ \frac{\gamma_i(\ln t_{pi} - (\beta_i - \tau_p))}{\alpha_i} \right]^2 \right\},
\]

with \( \tau_p \in \mathbb{R} \) representing the cognitive speed of test taker \( p \) and where \( \beta_i \in \mathbb{R} \) and
\( \gamma_i \in \mathbb{R}^+ \) are the parameters for the time intensity and discriminating power of
item \( i \), respectively. For technical details and applications of the model, we refer
to its introduction and comprehensive review in van der Linden (2006, 2016b).
Just as for the response model, the item parameters are assumed to be estimated
with enough precision during item calibration to treat them as known. As for the
estimation of the test takers’ speed parameters, it is already noted that, for a test
of \( I \) items, the maximum-likelihood estimates (MLEs) of these parameters have
the convenient closed form of

\[
\hat{\tau}_p = \left[ \sum_{i=1}^{I} \gamma_i^2 (\beta_i - \ln t_{pi}) \right] / \sum_{i=1}^{I} \gamma_i^2,
\]

which can be interpreted as the precision-weighted average of the differences
between the test taker’s logtimes and time intensities of the items. The asymp-
totic standard error of the MLEs is equal to
\[ \text{SE}(\hat{\tau}) = \left( \sum_{i=1}^{l} \alpha_i^2 \right)^{-1/2}, \]  

(van der Linden, 2016b, eqs. 16.13, 16.42). The consequences of using (4) to estimate the speed parameters of the test takers on the power of the RT-based test will be discussed after its presentation.

**Basic Hypotheses**

Let \( \gamma_i \in \{0, 1, \ldots, P\} \) be the number of test takers already familiar with item \( i \) prior to the test. This unknown parameter is equal to zero when the item has not been compromised but positive when it has. Thus, the two hypotheses that have to be tested against each other are

\[ H_0 : \gamma_i = 0, \]  

and

\[ H_1 : \gamma_i > 0. \]  

The two hypotheses need to be specified further using the parameterizations of the response model in (1)–(2) and RT model in (3).

**Response-Based Test**

**Hypotheses**

Assuming the response model holds for all regular test takers, the basic null hypothesis should be specified as

\[ H_0 : Pr\{U_{pi} = 1\} = \pi_{pi} \text{ for all test takers.} \]  

However, when a test taker already knew the item, the model is no longer valid. It then seems safe to assume that a test taker motivated to get hold of an item also has made sure to know the response, which implies a probability of observing a success equal to one. But, just for the sake of generality, as the proposed tests remain valid for any increase of \( \pi_{pi} \) due to item preknowledge, the alternative hypothesis is specified as

\[ H_1 : Pr\{U_{pi} = 1\} > \pi_{pi} \text{ for } \gamma_i \text{ of the test takers.} \]  

**Null and Alternative Distributions**

As the test takers are assumed to work independently during testing, the probability of observing a response vector \( u_i \equiv (u_{1i}, \ldots, u_{Pi}) \) follows from (1) as
\[ Pr\{U_i = u_i; P, \pi_{pi}\} = \prod_{p=1}^{P} \pi_{pi}^{u_p} (1 - \pi_{pi})^{1-u_p}. \]  
\[ (10) \]

Hence, the pmf of the total number of correct responses \( X_{pi} \) on the item is equal to

\[ f(x; P, \pi_i) = Pr\{X_{pi} = x; P, \pi_i\}, \]
\[ = \begin{cases} \sum_{u_p=0}^{x} \prod_{p=1}^{P} \pi_{pi}^{u_p} (1 - \pi_{pi})^{1-u_p}, & x = 0, 1, \ldots, P \\ 0, & \text{otherwise}, \end{cases} \]
\[ (11) \]

where \( \pi_i \equiv (\pi_{pi}, \ldots, \pi_{pi}) \). Distributions with this pmf are known to belong to the compound binomial family. Due to their combinatorial complexity, it may seem difficult to compute their probabilities, but, in fact, they are easily calculated using the well-known recursive algorithm introduced in the test-theory literature by Lord and Wingersky (1984).

In order to emphasize an analogy with the RT-based test of item compromise below, it is important to note that the algorithm consists of a sequence of convolution operations applied to the pmfs for the items in (1). For the case of \( P = 2 \), the operation is defined as

\[ (f_{2i} * f_{1i})(x) = \sum_{z=0}^{x} f_{2i}(x-z)f_{1i}(z), \]
\[ (12) \]

where the left-hand side is the pmf of the number of correct responses \( x = 0, 1, 2 \) on the item by the first two test takers. For the general case of \( P \) test takers, repeated application of the operation gives us the probabilities for their number of correct responses \( x = 0, 1, \ldots, P \) as

\[ f(x) = f_{pi} * (f_{(p-1)i} * \cdots * (f_{2i} * f_{1i}))(x). \]
\[ (13) \]

For a review of these operations and their applications in test theory, see van der Linden (2016a).

The alternative distribution is also compound binomial but this time with the success probabilities in (9). Let \( \Gamma_i \) be the set of \( \gamma_i \) test takers who knew the item in advance and \( \Gamma_i \) its complement. Formally, its pmf can then be written as

\[ f(x; P, \gamma_i, \pi_{\Gamma_i}) = \begin{cases} 0, & \text{for } x < \gamma_i, \\ f(x - \gamma_i; P - \gamma_i, \pi_{\Gamma_i}), & \text{for } \gamma_i \leq x < P, \end{cases} \]
\[ (14) \]

where \( \pi_{\Gamma_i} \) is the vector with the success probabilities for the regular test takers. This alternative distribution has zero probability for each of the \( x < \gamma_i \) successes in the window of \( P \) test takers, a simple consequence of the fact that at least \( \gamma_i \) of them already know the response. For the remaining part of its support, the distribution is a compound binomial over the \( P - \gamma_i \) regular test takers who did
not know the response in advance. This part is motivated by the fact that a total of exactly $g_i$ correct responses is observed when none of the regular test takers has the item correct, $g_i + 1$ if one of them has it correct, and so on.

Probabilities under the alternative distribution can only be calculated if the identities of the test takers who already knew the item are known. However, the fact that the distribution is from the same family as the null distribution is enough to establish the statistical test of item compromise in the next section.

**Statistical Test**

As shown by Romero et al. (2015), the family of compound binomial distributions has the property of a monotone likelihood ratio (MLR) in the number of successes $x$. Let $\pi_{0i} \equiv (\pi_{0Pi}, \ldots, \pi_{0Pi})$ and $\pi_{1i} \equiv (\pi_{1Pi}, \ldots, \pi_{1Pi})$ denote the success probabilities of the $P$ test takers on item $i$ under the null and alternative hypotheses in (8) and (9), respectively. Formally, this means that, for each vector inequality $\pi_{1i} > \pi_{0i}$, likelihood ratio

$$\frac{f(x; \pi_{1i})}{f(x; \pi_{0i})},$$

is a strictly increasing function of $x$. As follows from (9), the property immediately implies the MLR property for

$$\frac{f(x; \gamma_i > 0)}{f(x; \gamma_i = 0)},$$

as a function of $x$. We thus have a right-sided test of $H_0$ against $H_1$.

Let $x_{\text{crit}}$ be the smallest value of $X_{Pi}$ with right-tail probability $Pr\{X_{Pi} \geq x_{\text{crit}}|H_0\} \leq \alpha$ under the null hypothesis for a given significance level $\alpha$. The proposed test of $H_0: \gamma_i = 0$ against $H_1: \gamma_i > 0$ rejects the former in favor of the latter when

$$X_{Pi} \geq x_{\text{crit}}.$$  

Because of (16), according to the Neyman–Pearson lemma, a test of this type with known ability parameters $\theta_p$ is UMP (Casella & Berger, 2002, theorem 8.3.12). When planning an application of the test, it is important to know the power as a function of the possible values of $\gamma_i$ for the parameters of the items that are checked for compromise. As the power depends on the identity of the test takers in the subset with preknowledge of the item, rather than just its size, the best option is to calculate the functions for all possible subsets of test takers of the same size $\gamma_i$ and report their distribution. The process is illustrated in the empirical examples below. As the number of possible subsets quickly increased for values of $\gamma_i$ toward 0.5$P$, subsets of size $\gamma_i = 1, \ldots, P$ were randomly sampled from the set of simulated test takers. The sampling was replicated a large number of times to guarantee stable distributions of power functions.
The RT-based test starts from the same basic null and alternative hypotheses in (6)–(7), but this time they need to be translated into equivalent hypotheses about values for the parameters of the lognormal distributions of the RTs in (3) rather than the Bernoulli distributions in (1).

As the item is fixed but the test takers vary, an obvious choice is to focus on the test takers’ speed parameters. More importantly, the choice makes substantive sense too: Test takers who know the item in advance can be expected to respond faster than when they actually need to read, understand, and solve it in real time. Let \( \tau_p \) denote the speed parameter on the tests for a regular test taker without preknowledge of any of the items. For the case of a test based on RTs, the two basic hypotheses in (6)–(7) then specialize to

\[
H_0 : \tau_{pi} = \tau_p \text{ for all test takers},
\]

\[
H_1 : \tau_{pi} > \tau_p \text{ for } \gamma_i \text{ of the test takers.}
\]

The null distribution now changes from one of the sums of discrete responses, \( X_{pi} \), to sums of continuous RTs, \( T_{pi} \). Still assuming test takers who have worked independently of each other, it follows we are required to use convolution integrals rather than sums to calculate the distribution of \( T_{pi} \) under the two hypotheses. That is, for the case of two examinees \( p = 1 \) and two with pdfs \( f_{1i} \) and \( f_{2i} \) for their RT distributions on item \( i \), the sum of their two RTs has pdf

\[
(f_{2i} * f_{1i})(t) = \int_0^t f_{2i}(z)f_{1i}(t-z)dz.
\]

Repeated convolution of the result with the pdf of an additional test taker, similar to (13), gives us the distribution of \( T_{pi} \) for the \( P \) test takers. However, just as for the Bernoulli distributions, the family of lognormal distributions is not known to be closed under convolution. To make things worse, a recursive algorithm analogous to Lord and Wingersky’s is practically infeasible. The algorithm would have to evaluate the integrals present in each of the \( P - 1 \) steps in (13) numerically for real values of \( t \), which is just an impossible task. However, a convenient way around the obstacle is possible capitalizing on the fact that the pdf of the lognormal distribution of a random variable actually is another representation of the normal pdf of its logarithm. The normal family is known to be closed under the convolution operation. As a bonus, a well-known property of
the family also allows us to precisely pinpoint the statistical nature of the desired statistical test.

Let

\[ T^*_{pi} = \ln T_{pi}, \]  

(21)

and define

\[ T^*_p \equiv \sum_{p=1}^{P} T^*_pi, \]

(22)

as their sum across all \( P \) test takers. For the parameterization in (3), the null distribution of \( T^*_p \) has the normal pdf

\[ f(t^*; \mu_{Pi}, \sigma_{Pi}) = \frac{1}{\sigma_{Pi} \sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left( \frac{t^* - \mu_{Pi}}{\sigma_{Pi}} \right)^2 \right\}, \]

(23)

with mean and variance equal to

\[ \mu_{Pi} = P\beta_i - \sum_{p=1}^{P} \tau_{pi}, \]

(24)

and

\[ \sigma^2_{Pi} = P\varsigma_i^{-2}. \]

(25)

The alternative distribution is a normal with the same variance but a mean equal to

\[ \mu_{Pi} = P\beta_i - \sum_{p=1}^{P} \tau_{pi}, \]

(26)

that is, with the sum of the actual speed parameters for the \( P \) test takers on the item. The sum is larger due to the increase of speed for each of the test takers with preknowledge.

\textit{Statistical Test}

As \( P \) is known by design and item parameters, \( \alpha_i \) and \( \beta_i \) are assumed to be estimated with enough precision during item calibration to be treated as known, both the null and alternative distribution belong to the family of normal distributions with known variance but unknown mean. The family has the MLR property in a sufficient statistic for its mean (Casella & Berger, 2002, example 8.3.15). Because (24) decreases with an increase of \( \tau_{pi} \) for each of the test takers, the property also holds for the ratio of

\[ \frac{f(t^*; \gamma_i > 0)}{f(t^*; \gamma_i = 0)}. \]

(27)
as a function of $t^*$, which implies a left-sided test of $H_0$ against $H_1$. Using $z_a$ to denote the $a$th quantile in the standard normal distribution, we therefore reject $H_0 : \gamma_i = 0$ in favor of $H_1 : \gamma_i > 0$ when $(T_{pi}^* - \mu_{pi}) / \sigma_{pi} < z_a$. Or, in terms of the RT parameters, when the sum of the logtimes, $T_{pi}^*$, is smaller than $t_{crit}' = P\beta_i - \sum_{p=1}^{P} \tau_p + z_aP^{1/2} \alpha_i^{-1}$. (28)

Observe the simplicity of the calculations required to apply the RT-based test during operational testing. The only data that need to be collected are the log-RTs of the test takers along with the estimates of their speed parameters. The item is then flagged for potential compromise when their sum $\sum_{p=1}^{P} (T_{pi}^* + \tau_p)$, (29) is smaller than the constant $P\beta_i + z_aP^{1/2} \alpha_i^{-1}$. (30)

Just as the response-based test, because of its MLR property, the test is also UMP for known speed parameters $\tau_p$.

**Examples of Power Functions**

Examples of the power functions for both types of tests are given. The examples are given for the case of known ability and speed parameters. The impact of estimation of these parameters is discussed immediately after their presentation.

As the response-based test in (17) is right sided, its power function is probability

$Pr\{X_{pi} \geq x_{crit}|\gamma_i\}$, (31)

for the family of compound binomial distributions as a function of $\gamma_i = 0, 1, ..., P$. The function is illustrated for items with the four combinations of difficulty parameters $b_i = -1.0$ and $b_i = 1.0$, discrimination parameters $a_i = 0.6$ and $a = 1.4$, and common guessing parameter $c_i = 0.25$. The ability parameters of $P = 50$ test takers were randomly sampled from $N(0, 1)$, a distribution assumed to be imposed as identification constraint during item calibration. Together, the four combinations should cover the ranges of parameter values for four-choice items in a well-designed pool for the assumed population of test takers.

To calculate the functions, the following steps were taken: The response probabilities were calculated for the sample of 50 test takers from the population distribution. Each of the subsets $\Gamma_i$, $\gamma_i = 1, ..., P - 1$, was sampled 1,000 times from the 50 test takers, a sample size assumed to be large enough to estimate the power for each value of $\gamma_i$ with desirable accuracy. Assuming test takers who
know compromised items along with their answers, probabilities of success for the test takers in these random subsets were adjusted setting them equal to $p_i = 1$. For each replication, the compound binomial distribution was calculated using the Lord–Wingersky algorithm and the probability in (31) was determined for the critical value $x_{\text{crit}}$ in (17) at significance level $\alpha = .05$. Figure 1 shows the 5th, 25th, 50th, 75th, and 95th quantiles of the distribution of these probabilities across the 1,000 replications as a function of the number of test takers with preknowledge of the items, $\gamma_i = 0, 1, \ldots, 50$.

The results point at a statistical test that detects item compromise with nearly perfect power when approximately 20 test takers have preknowledge of one of

FIGURE 1. Quantiles of the power distributions of the response-based test of item compromise as function of the number of test takers with preknowledge of items with four different combinations of difficulty and discrimination parameters. Curves from the right to the left are for the 5th, 25th, 50th, 75th, and 95th quantiles, respectively.
The two more difficult items, whereas some 30 were necessary for the two easier
items. As the sets $\Gamma_i$ were randomly sampled, the conclusion holds independently
of the order in which such test takers would show up. The results also revealed a
general trend across the different combinations of item parameter values: The
power functions appear to be steeper, reach the maximum power of one for a
smaller number of test takers with preknowledge, and show less variation across
the sampled subsets $\Gamma_i$ for more difficult items. The trend can be explained by the
fact that, for such items, when moving up the ability scale, their regular prob-
ability of success approaches the value of one later than for the less difficult
items. As a consequence, they give the statistical test more power to discriminate
between the probabilities of success for the regular and cheating test takers across
a greater segment of the scale.

The RT-based test is left sided and its power is defined as

$$Pr\{T_{Pi} \leq t_{\text{crit}} | \gamma_i\},$$

for the family of normal distributions in (23)–(26) as a function of $\gamma_i = 0, \ldots, P$.
Two features of the family relevant to the nature of the power functions should be
noticed. The first is the fact that the size of the overlap between the null and
alternative distributions only depends on discrimination parameter $\alpha_i$ and the
sum of the actual speed parameters of the test takers on the item. Specifically, it is
completely independent of time-intensity parameter $\beta_i$. A change in the size of
this parameter introduces the same shift in the location of the null distribution as
for each of the alternative distributions. As the power of the test is a measure of
the overlap between the two, we thus always have identical power functions for
the RT-based test for items with different time intensities but the same discri-
minating power. The second feature is that, to calculate the power of the test, it is
no longer necessary to know the identity of the test takers in subset $\Gamma_i$ with the
preknowledge of the item. In fact, it is not even necessary to know its size. The
only thing that matters is the sum of their individual increases in speed due to
prior knowledge of the item.

To illustrate the shape of the power functions, four items with common
arbitrary time-intensity parameter $\beta_i = 4.0$, but different discrimination para-

meters $\alpha_i = 2.3, 2.0, 1.7, \text{ and } 1.4$ were used. The values for the item parameters
were typical of the items in the empirical data set in van der Linden (2006), which
had a range of [3.14, 4.91] for their estimated time-intensity and [1.38, 2.31] for
their estimated discrimination parameters. For the same data set, the estimated
speed parameters had a mean of zero as the result of an identification constraint
and a standard deviation $\sigma_\tau = 0.35$. The functions were calculated as follows:
The speed parameters of $P = 50$ test takers were randomly sampled from
$N(0, 0.35^2)$. Just as for the response-based test, each of the subsets $\Gamma_i,$
$\gamma_i = 1, \ldots, P - 1,$ was sampled 1,000 times from the 50 test takers. The choice
of a realistic increase of the speed parameters for the test takers with preknow-
ledge of the item is important. Recently, Zopluoglu et al. (2021) reported the
results from an experiment, in which they compared the speed of test takers on a portion of the GRE Quantitative Reasoning Test with some of the items and answer keys disclosed to the test takers prior to the test relative to the control condition of nothing disclosed at all. The total set of items had a range of $[3.51, 2.16]$ for the time-intensity and $[0.83, 2.16]$ for the discrimination parameters estimated under a version of the lognormal RT model extended with an indicator variable for the disclosed versus nondisclosed items. The standard deviation of the distribution of the speed parameters was estimated as $\sigma_t = 0.38$. The average increase in speed for the test takers on the items known in advance appeared to be 1.65. With the exception of a somewhat smaller size for the discrimination parameters, the ranges of the parameters in this study are close enough to the estimates in our data set to accept the increase as a point of departure. Four alternative increases were adopted to serve for the calculation of the power functions in our example: 1.85, 1.65, 1.45, and 1.25. The addition of the smallest increase was to produce an extra example on the more conservative side.

Figure 2 shows the average power functions for critical value $t_{\text{crit}}$ in (28) at significance level $\alpha = .05$ as a function of $\gamma_i$. The functions appear to be generally steep and pointing at high power. For each of the combinations of level of discrimination and increase in speed, no more than some 20 test takers with preknowledge of the item was required to detect its compromise with power close to one. For the item with the highest discrimination parameter, some 10 test takers were already sufficient. These results tend to be even more promising than for the response-based test. Also, remember that they generalize to items of any time intensity. Finally, notice that, just as for Figure 1, the power of the test for $\gamma_i = 0$ is equal to 0.05, which was the nominal level of significance adopted for the two tests.

**Estimation of Ability and Speed Parameters**

When using the two proposed tests in real-world testing, the ability and speed parameters should be estimated from the test takers’ responses and RTs on the regular items in the test. A common solution in this kind of situation is use of the well-known leave-one-out method. The method should be used recursively with a first stage with each monitored item removed from estimation at a time and subsequent stages, where all items found to be suspicious are removed from the estimation until stability is obtained.

Speed parameter $\tau_p$ in the RT-based test can be estimated using the MLE and standard error with their simple expressions in (4) and (5). As an example, the standard error is already as low as .01 for an estimate of the parameter from 25 regular items with discrimination parameters $\alpha_i$ equal to 1.85 (−midpoint of the range for the items in Figure 2). Remember that the standard error is independent of the true value of $\tau_p$. The example is thus valid for every test taker, no matter how
The proposed statistical test is sensitive to the sum of the speed parameter estimates for the \( P \) test takers in (24). If the test takers can be assumed not to have communicated during the test, we have a sum of estimates with independent errors in its terms and a standard error increasing with the square root of \( P \). Thus, continuing the example, for a window size of \( P = 25 \) test takers, which was sufficient to detect compromise with complete power for the cases in Figure 2, the standard error of the sum could be expected to be as low as .05, a value suggesting minor impact on the statistical test.

Software for the estimation of the ability parameters from the test takers’ responses on calibrated items during operational testing is widely in use in the

FIGURE 2. Average power functions of the response time–based test of item compromise for items with four different levels of discrimination, arbitrary time intensity, and four alternative increases of the test takers’ speed due to item preknowledge. Curves from the left to the right are for an increase of speed equal to 1.85, 1.65, 1.45, and 1.25, respectively.
Two Statistical Tests for the Detection of Item Compromise

testing industry. More efficient estimation of both types of test taker parameters is even possible estimating them jointly using hierarchical modeling (van der Linden, 2007; C. Wang & Xu, 2015). Although the above argument should also hold for the response-based test, we refrain from speculation about the quantitative size of the impact of estimation of the ability parameters on its power because unlike the RT-based test, closed-form expressions as in (4) are not available for the response model in (2). The standard error of estimation does depend on the true value of the ability parameter, but, for a well-designed test, the ability and item parameters match and the error is minimal. It is recommended to evaluate the impact empirically when the item parameters and the expected range of ability parameters are known prior to the use of the test in operational testing.

It is important to note that loss of power due to parameter estimation does not necessarily imply loss of optimality for a statistical test relative to the class of tests requiring estimation of the same parameters. A necessary condition for maintaining optimality is estimation based on a sufficient statistic, a condition met for MLEs of the individual speed and ability parameters. The invariance property of MLE guarantees the same condition to be met for the sum of the speed parameters and the vector of success probabilities serving as parameters of interest in the two tests. However, observe that we are not faced with the standard situation of a statistical test of a hypothesis about an intentional parameter with a null distribution requiring the estimation of one or more nuisance parameters (as, for instance, when the variance of the distribution for the RT-based test would have been unknown). Also, the test statistics for the two proposed tests of item compromise and the speed and ability estimates are calculated from the responses and RTs collected from different sets of items. Having been unable to find theory about this specific case, the author refrains from conclusions about sufficient conditions for optimality of the two proposed tests.

Discussion

An issue still investigated is how to use the two proposed tests in combination. The standard assumption of conditional (or “local”) independence between responses and RTs on each item, which, under the null hypothesis of \( \gamma_i = 0 \), underlies the use of the 3PL and lognormal models in this paper, suggests gain of power relative to the individual tests. The relevant question, of course, is how to combine them. One option is to simply flag an item as compromised as soon as either test is significant. An obvious alternative is to require both tests to be significant before raising the flag. A third alternative is to apply one of the two tests at the level of each single test taker and use the other test capitalizing on the outcomes from the first. Initial results for the last alternative suggest an interesting gain of power relative to individual use of either test, but more research is needed to confirm the generality of the conclusion. For example, as the responses and RTs are now used jointly, extensive evaluation of the impact of the
correlation between the speed and abilities of the test takers on the power of the possible combinations is required. Final results will be reported in van der Linden and Belov (2022).

The natural follow-up of a test of item compromise is an attempt to identify the test takers who actually profited from the compromised item. The options for doing so depend on the number of items monitored simultaneously. If an incidental item is monitored, one response and RT per test taker is available to check on their integrity. The only possible option then seems residual analysis, looking for test takers with a combination of low ability and speed producing a correct response in a short time on a more difficult, time-intensive item, as demonstrated in the empirical example in van der Linden and Guo (2008). The option is less effective for compromised items that are relatively easy and/or less time intensive though. Better practice seems to monitor all items in the pool simultaneously and follow-up with one of the statistical tests of item preknowledge for the identified set of compromised items reviewed in the introduction to this article, for example, the likelihood-ratio test by Sinharay and Johnson (2020).

The issue of power deserves not only statistical attention though. Ultimately, the combination of the choice of window size and significance level for a test of item compromise should depend on the evaluation of the relative consequences of a false positive and false negative decision to flag an item as suspicious. If a testing program has enough pretested items in stock to replace a suspicious item, a false negative decision will do much more harm to the program than replacing it too early. It could even be argued that, for such programs, the significance level should be set higher than the traditional choice of $\alpha = .05$ to obtain more even power than demonstrated by the empirical examples presented in this article.

Finally, it is important to remember that, no matter its power, a significant result for a statistical test only satisfies a necessary condition for the alternative hypothesis to be true. For example, higher actual speed on an item than expected for a test taker could also be the result of guessing on items located toward the end of speeded test. Lack of sufficient evidence is probably less of a concern to a testing program willing to accept a false positive decision as price for a conservative policy. However, it definitely should be a concern when follow-up statistical tests are used to check on test takers with preknowledge on items flagged as compromised. It is then always important to supplement the tests with evidence in the form of reports of observed irregular behavior during testing, detected visits to websites offering stolen items prior to the test, or membership of cliques of test takers with possible preknowledge of the same set of items (e.g., Belov & Wollack, 2021).

Author’s Note
The opinions and conclusions contained in this article are those of the author and do not necessarily reflect the policy and position of the Law Schools Admission Council.
Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research and/or authorship of this article: This study received funding from the Law Schools Admission Council.

References

Belov, D. I. (2014). Detecting item preknowledge in computerized adaptive testing using information theory and combinatorial optimization. *Journal of Computerized Adaptive Testing, 2*(3), 37–58.
Belov, D. I. (2016). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement, 40*, 83–97.
Belov, D. I., & Armstrong, R. D. (2011). Distributions of the Kullback–Leibner divergence with applications. *British Journal of Mathematical and Statistical Psychology, 64*, 291–309.
Belov, D. I., & Wollack, J. A. (2021). Graph theory approach to detect examinees involved in test collusion. *Applied Psychological Measurement, 45*, 253–267.
Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Duxbury.
Cheng, Y., & Yuan, K.-H. (2010). The impact of fallible item parameter estimates on latent trait recovery. *Psychometrika, 75*, 280–291.
Liu, Y, & Yang, J. S. (2018). Bootstrap-calibrated interval estimates for latent variable scores in item response theory. *Psychometrika, 83*, 333–354.
Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile equating. *Applied Psychological Measurement, 8*, 453–461.
Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics, 39*, 426–451.
McLeod, L., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement, 27*, 121–137.
Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensing examinations. *Educational Measurement: Issues and Practice, 35*, 38–47.
Romero, M., Riascos, Á., & Jara, D. (2015). On the optimality of answer-copying indices: Theory and practice. *Journal of Educational and Behavioral Statistics, 40*, 435–453 (Corrigendum, 2016, 41, 659).
Segall, D. O. (2002). An item response model for characterizing test compromise. *Journal of Educational and Behavioral Statistics, 27*, 163–179.
Segall, D. O. (2004). A sharing item response theory model for computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 29*, 439–460.
Sinharay, S. (2016). Person fit analysis in computerized adaptive testing using tests for a change point. *Journal of Educational and Behavioral Statistics, 41*, 521–549.
van der Linden

Sinharay, S. (2017a). Some remarks on applications of tests for detecting a change point to psychometric problems. *Psychometrika*, 82, 1149–1161.

Sinharay, S. (2017b). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics*, 42, 46–68.

Sinharay, S., & Johnson, M. S. (2020). The use of item scores and response times to detect examinees who may have profited from item preknowledge. *British Journal of Mathematical and Statistical Psychology*, 73, 397–419.

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. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72, 287–308.

van der Linden, W. J. (2016a). Distributions of sums of nonidentical random variables. In W. J. van der Linden (Ed.), *Handbook of item response theory: Volume 2. Statistical tools* (pp. 97–103). Chapman & Hall/CRC.

van der Linden, W. J. (2016b). Lognormal response-time model. In W. J. van der Linden (Ed.), *Handbook of item response theory: Volume 1. Models* (pp. 261–282). Chapman & Hall/CRC.

van der Linden, W. J., & Belov, D. (2022). *A statistical test of item compromise based on a combination of responses and response times* [Manuscript to be submitted for publication]. Department of Behavioral, Management and Social Sciences, University of Twente.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, 73, 287–308.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics*, 26, 199–218.

Veerkamp, W. J. J., & Glas, C. A. W. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics*, 25, 373–389.

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.

Wang, X., & Liu, Y. (2020). Detecting compromised items using information from secure items. *Journal of Educational and Behavioral Statistics*, 45, 667–689.

Wang, X., Liu, Y., & Hambleton, R. K. (2017). Detecting item preknowledge using a predictive checking method. *Applied Psychological Measurement*, 41, 243–263.

Yang, J. S., Hansen, M., & Cai, L. (2012). Characterizing sources of uncertainty in item response theory scale scores. *Educational and Psychological Measurement*, 72, 264–290.

Zhang, J. (2014). A sequential procedure for detecting compromised items in the item pool of a CAT system. *Applied Psychological Measurement*, 38, 87–104.

Zhang, J., & Li, J. (2016). Monitoring items in real time to enhance CAT security. *Journal of Educational Measurement*, 53, 131–151.

Zopluoglu, C. Z., Kasli, M., & Toton, S. L. (2021). The effect of item preknowledge on response time: An analysis of two data sets using the multi-group lognormal response model with gating mechanism. *Educational Measurement: Issues and Practice*, 40, 42–51.
Two Statistical Tests for the Detection of Item Compromise

Author

WIM J. VAN DER LINDEN is professor emeritus of measurement and data analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands; email: wjvdlinden@outlook.com. His interests include test theory, applied statistics, and research methods.

Manuscript received May 14, 2021
Revision received March 8, 2022
Accepted March 23, 2022