JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education
DOI: 10.1111/jedm.12346 *Journal of Educational Measurement*

A Statistical Test for the Detection of Item Compromise Combining Responses and Response Times

Wim J. van der Linden
University of Twente
Dmitry I. Belov
Law School Admission Council

A test of item compromise is presented which combines the test takers' responses and response times (RTs) into a statistic defined as the number of correct responses on the item for test takers with RTs flagged as suspicious. The test has null and alternative distributions belonging to the well-known family of compound binomial distributions, is simple to calculate, and has results that are easy to interpret. It also demonstrated nearly perfect power for the detection of compromise with no more than 10 test takers with preknowledge of the more difficult and discriminating items in a set of empirical examples. For the easier and less discriminating items, the presence of some 20 test takers with preknowledge still sufficed. A test based on the reverse statistic of the total time by test takers with responses flagged as suspicious may seem a natural alternative but misses the property of a monotone likelihood ratio necessary to decide between a test that should be left or right sided.

Introduction

Cheating on tests has been on the rise for some time now, particularly for high-stakes tests used for admission to educational programs and licensing for professions with limited access. Though widely spread, the problem is particularly serious for programs that offer online continuous testing with individual candidates signing up for administration on their day of choice. These programs typically are well protected against attempts to copy answers from fellow test takers during the test but have become subject to attempts to harvest items from their pools which are then shared, or even sold for financial gain, to future test takers. Testing programs have been on the alert of the development and implemented lines of defense as frequent item pool refreshment, randomized item selection techniques to control the exposure rates of their items, and the use of data forensic techniques. The statistical test of item compromise proposed in this paper belongs to the last category.

Earlier statistical tests to detect cheating have been based exclusively on statistics defined on the responses or response times (RTs) on the items. Examples of tests based on responses include the use of the cumulative sum (CUSUM) technique from statistical quality control by Veerkamp and Glas (2000) and change-point analysis by Chen, Lee, and Li (2021), Sinharay (2016, 2017a), Zhang (2014), and Zhang and Li (2016). Bayesian versions of response-based tests have been proposed by Belov (2016), Belov and Armstrong (2011), McLeod, Lewis, and Thissen (2003),

and Wang, Liu, and Hambleton (2017). The idea to detect cheating using RTs rather than responses has been explored, for instance, by Marianti, Fox, Avetisyan, and Veldkamp (2014), Qian, Staniewska, Reckase, and Woo (2016), Sinharay (2020), Wang and Liu (2020), and van der Linden and Guo (2008). For examples of in-depth reviews of several of these tests, see Belov (2016), Sinharay (2017b), and van der Linden (2018).

Earlier attempts to combine responses and RTs in one analysis include Choe and Chang (2014), Choe, Zhang, and Chang (2018), and van der Linden and Guo (2008). The current research was particularly inspired by Belov and Cubbellotti (2017). The focus of their method was on the detection of suspicious changes of responses on answer sheets with as focal quantity the average shift in the posterior distributions of the test taker's ability and speed parameter given the responses and RTs on the items with changed and unchanged responses. Recently, Sinharay and Johnson (2020) proposed a constrained likelihood-ratio statistic based both on responses and RTs. The statistic can be used to test the joint hypothesis of the test takers' parameters in the hierarchical model of speed and accuracy (van der Linden, 2007) being the same on the uncompromised items in the test and items suspected to be compromised. The Sinharay and Johnson test is thus focused on the detection of test takers with preknowledge rather than compromised items. The same idea of combining responses and RTs to detect cheating was used by Belov and Toton (2022) in a study to detect cliques of test takers with preknowledge. Once a clique is detected, items with common answers across the clique can be used to detect the subset of compromised items they have shared (Belov & Wollack, 2021).

The same idea of combining responses and RTs to detect item compromise is investigated in the current study. It does not assume anything known about these items or the test takers other than estimates of their regular parameters. Once compromised items have been detected, it is natural to follow up with one of the statistical tests available to identify the test takers who might have profited from preknowledge of them, for example, the one introduced by Drasgow, Levine, & Zickar (1996). Though it is certainly possible to use the test as part of forensic analysis after a group-based fixed form has been administered, it is presented for the case of real-time monitoring of test items for possible compromise in a continuous online testing program. The test offered by the program may have any adaptive or fixed format. The only thing necessary is the collection of the responses and RTs for a window of test takers during the active stage of the program.

The first section below introduces examples of the response and RT models for which the proposed test could be used and presents the null and alternative hypotheses on an unknown parameter representing item compromise that need to be tested against each other to detect compromise. The following section introduces the test statistic defined as the sum of the responses for test takers in the window with RTs flagged as suspicious and identifies the family to which its null and alternative distributions belong. The parameters for these distributions are derived from a recent RT-based test applied at the level of the individual test takers and a response-based test at the level of the entire window. Both separate tests are briefly reviewed in the Appendix. Next, examples of the power functions for a set of realistic choices of item parameters for the response and RT model are presented. In the last part of the paper,

a seemingly obvious alternative test based on the total time by the test takers with responses identified as suspicious is explored and its unexpected behavior is empirically demonstrated. Finally, a few remaining issues related to practical application of the test are discussed.

Basic Setup

Suppose a set of test items is monitored to determine if some of them have been compromised. For an item i, the observation consists of the random responses U_{pi} and RTs T_{pi} collected for a window of $p = 1, \ldots, P$ test takers. The monitoring could be continuous, in a periodic manner, or after the test has been administered for an extended period. The choice of format and size of the window are up to the testing program. As will appear below, it is convenient to work with $T^*_{pi} = \ln T_{pi}$ rather than the RTs recorded in their standard metric.

The distribution of U_{pi} is Bernoulli with probability mass function (pmf)

$$f(u_{pi}; \pi_{pi}) = \pi_{pi}^{u_{pi}}(1 - \pi_{pi})^{1-u_{pi}}, \tag{1}$$

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

$$\pi_{pi} \equiv c_i + (1 - c_i)\big[1 + \exp(-a_i(\theta_p - b_i))\big]^{-1}, \tag{2}$$

with θ_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 adopted to account for the effect of guessing. The item parameters in the model are supposed to have been estimated during earlier item calibration with enough precision to treat them as known. The ability parameters of the test takers are assumed to be estimated from the items in the test by one of the standard software programs in use for the model. If there exists evidence that an item may be compromised, the parameters could be estimated just from the other, potentially uncompromised items. But if the statistical test is used to monitor all items in the pool, a common approach is to estimate them using the leave-one-out method well-known from its use in cross-validation studies (e.g., Hastie, Tibshirani, & Friedman, 2009, chap. 7). The method removes each monitored item from the estimation at a time and then follows up with recursive steps in which all items flagged as compromised at the preceding step are removed from the estimation until stability is obtained. The choice of the 3PL model is not restrictive. Any model with separate parameters for the ability of the test takers and properties of the item that explains the probabilities could have been chosen to derive the statistical test below.

The model for the RTs is the lognormal model which postulates the distribution of the RTs T_{pi} to have probability density function (pdf)

$$f(t_{pi}) \equiv \frac{\alpha_i}{t_{pi}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\big[\alpha_i(\ln t_{pi} - (\beta_i - \tau_p))\big]^2\right\}, \tag{3}$$

where $\tau_p \in \mathbb{R}$ is interpreted as the cognitive speed of test taker p and $\beta_i \in \mathbb{R}$ and $\alpha_i \in \mathbb{R}^+$ are parameters for the time intensity and discriminating power of item i, respectively. The pdf of the lognormal distribution of a random variable actually is an alternative representation of the normal pdf for its logarithm. Thus, for the parameterization in (3), we can use

$$f(t_{pi}^*; \mu_{Pi}, \sigma_{Pi}) = \frac{1}{\sigma_{pi}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{t^* - \mu_{pi}}{\sigma_{pi}}\right)^2\right\} \tag{4}$$

with mean and variance equal to

$$\mu_{pi} = \beta_i - \tau_p \tag{5}$$

and

$$\sigma_{pi} = \alpha_i^{-1}. \tag{6}$$

Just as for the response model, the item parameters are assumed to have been estimated with enough precision during item calibration to treat them as known while the test takers' speed parameters are estimated in a similar fashion along with the ability parameters. Estimation of the speed parameters is straightforward as, for a set of I items, the maximum-likelihood estimator (MLE) of the speed parameters has the simple expression of

$$\hat{\tau}_p = \left[\sum_{i=1}^{I} \alpha_i^2 (\beta_i - \ln t_{pi})\right] \Bigg/ \sum_{i=1}^{I} \alpha_i^2 \ , \tag{7}$$

with an asymptotic standard error equal to

$$\mathrm{SE}(\hat{\tau}) = \left(\sum_{i=1}^{I} \alpha_i^2\right)^{-1/2} . \tag{8}$$

Further technical details and applications of the model are provided in van der Linden (2016a).

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

$$H_0 : \gamma_i = 0 \tag{9}$$

and

$$H_1 : \gamma_i > 0. \tag{10}$$

The hypotheses can be tested using a statistical test based exclusively on the responses or the RTs for the test takers. However, under the null hypothesis, as the responses and RTs of the test takers can be assumed to be two conditionally ("locally") independent sources of information given the ability and speed of the test takers, a natural question is if it is possible to improve on the power of separate tests by combining the two.

A Combined Test of Item Compromise

The idea is investigated in the current study combining the two separate tests of item compromise in van der Linden (2022), which are briefly reviewed in the Appendix. The response-based test has as test statistic the sum of the responses

$$X_{Pi} = \sum_{p=1}^{P} U_{pi}, \tag{11}$$

while the sum of the log-RTs,

$$T^*_{Pi} = \sum_{p=1}^{P} T^*_{pi}, \tag{12}$$

serves as statistic for the RT-based test.

The proposed new test is based on a combination of the two statistics defined as

$$Z_{Pi} = \sum_{p=1}^{P} Q_{pi} U_{pi}, \tag{13}$$

where $Q_{pi} \in \{0, 1\}$ is an indicator variable that takes the value of one for test taker p with an RT on item i flagged as significant and zero otherwise, a decision based on the lower tail of the item-level RT distribution in (4). One could think of Q_{pi} as a time-based weight for the responses indicating whether or not the test taker has produced a suspiciously short RT. Our hypothesis is that, though based on a smaller number of responses, thanks to the additional information provided by the RTs, a test with statistic Z_{Pi} may be more informative than one based on the plain sum of all responses.

Empirical support for the hypothesis was particularly suggested by the study of the clique detector for the detection of test takers with preknowledge in Belov and Toton (2022). When these authors used a similarity index combining responses and RTs, they detected twice as many of the cheaters in the well-known empirical data set in Toton and Maynes (2019) than for an index derived from the Drasgow et al. (1996) test based on responses only. The best way to determine if such increased effectiveness holds generally is to evaluate the impact on the power function of the test.

As $Z_{pi} = Q_{pi}U_{pi} \in \{0, 1\}$ for each of the P test takers, Z_{Pi} still is the result of a sequence of Bernoulli trials with unequal success probabilities. The new statistic thus also has null and alternative distributions belonging to the compound binomial family (van der Linden, 2016b). Indicator variable Q_{pi} is a random variable which, for each $p = 1, \ldots, P$, has pmf

$$f(q_{pi}) = \begin{cases} 1 - F(t^*_{\text{crit}}; \tau_p, \alpha_i, \beta_i), & \text{for } q_{pi} = 0, \\ F(t^*_{\text{crit}}; \tau_p, \alpha_i, \beta_i), & \text{for } q_{pi} = 1, \end{cases} \tag{14}$$

where $F(\cdot)$ is the cumulative density function (cdf) for the log-RT distribution in (4)–(6) and

$$t^*_{\text{crit}} = \beta - \tau_p + z_a \alpha_i^{-1} \tag{15}$$

is the critical value for the version of the RT-based test in (A.11) applied at the level of a single test taker. As already indicated in (1), response variable U_{pi} has pmf

$$f(u_{pi}) = \begin{cases} 1 - \pi_{pi}, & \text{for } u_{pi} = 0, \\ \pi_{pi}, & \text{for } u_{pi} = 1. \end{cases} \tag{16}$$

The event of $Z_{pi} = 1$ occurs when both $Q_{pi} = 1$ and $U_{pi} = 1$. Thus, because of the conditional independence, Z_{pi} has pmf

$$f(z_{pi}) = \begin{cases} 1 - \pi_{pi} * F(t^*_{\text{crit}}; \tau_p, \alpha_i, \beta_i), & \text{for } z_{pi} = 0, \\ \pi_{pi} * F(t^*_{\text{crit}}; \tau_p, \alpha_i, \beta_i), & \text{for } z_{pi} = 1. \end{cases} \tag{17}$$

The null distribution is the member of the compound binomial family with success probabilities determined by the probabilities for the regular responses and RTs in (17). The alternative distributions have γ_i of these success probabilities inflated due to preknowledge of the item. Observe that the family still enjoys the property of a monotone likelihood ratio (MLR) observed for the separate response-based test in the Appendix. For each additional test taker with preknowledge of the item, the probability of $Z_{pi} = 1$ increases due to an increase both of success probability π_{pi} and speed parameter τ_p. The test thus remains right-sided with a critical value as in (A.5).

Examples of Power Functions

The power function for the proposed test is

$$\Pr\{Z_{Pi} \geq z_{\text{crit}} \mid \gamma_i\} \tag{18}$$

as a function of $\gamma_i = 0, 1, \ldots, P$. The function is important in that it summarizes all statistical properties of the test; for instance, note that, for $\gamma_i = 0$, it returns the actual significance level of the test. Also, observe that the function follows directly from the family of compound binomial distributions with as parameters the probabilities of success in (17). Particularly, it is not necessary to simulate any responses or RTs from their distributions in (1)–(2) and (3).

The functions are illustrated for items with the same four combinations of difficulty parameters $b_i = -1.0$ and 1.0, discrimination parameters $a_i = .6$ and 1.4, and common guessing parameter $c_i = .25$ as for the examples for the separate response-based test in van der Linden (2022). As explained in this reference, though the power for the RT-based test does depend on speed parameters τ_p and discrimination parameter a_i, it appears to be completely independent of time-intensity parameter β_i. The power functions are therefore demonstrated for items with $\alpha_i = 2.3$ and 1.4, but common arbitrary choice of $\beta_i = 4.0$ for each of the items in the set of examples. In addition, to assess the impact of possible correlation between the test takers' ability and speed, two extra conditions with $\rho_{\theta\tau} = .0$ and .6 were introduced.

The steps taken to calculate the functions were as follows:

1. $P = 50$ test takers were sampled from $\theta \sim N(0, 1^2)$ and their regular probabilities of success π_{pi} for the 3PL model were calculated.

2. Assuming bivariate normality, the speed parameters for these test takers were sampled from their conditional distributions given θ, using

$$\tau_p|\theta_p \sim N(\mu_\tau + \rho_{\theta\tau}\frac{\sigma_\tau}{\sigma_\theta}(\theta_p - \mu_\theta), \sigma_\tau^2(1 - \rho_{\theta\tau}^2)), \tag{19}$$

   where μ_τ and σ_τ^2 were equal to 0 and .35, respectively. To avoid correlations higher or lower than intended due to randomness, the sampling was repeated until two vectors of length $P = 50$ with empirical correlation $r_{\theta\tau}$ negligibly close to the intended values of .0 and .6 were obtained.
3. The cases of $\gamma_i = 1, \ldots, 50$ test takers with preknowledge were randomly picked from the array of simulated test takers.
4. The null and alternative distributions were calculated using the Lord-Wingersky algorithm for the compound binomial distribution with as parameters the probabilities of $z_{pi} = 0$ and $z_{pi} = 1$ in (17). For the regular test takers, π_{pi} and $F(t^*_{\text{crit}})$ were calculated from the ability and speed parameters sampled in Steps 1 and 2, with the level of significance for the RT-based test set at $\alpha = .05$. However, for the test takers with preknowledge, assuming they know the correct answer as well, the success probabilities were set equal to $\pi_{pi} = 1$ while $F(t^*_{\text{crit}}; \tau_p, \alpha_i, \beta_i)$ was calculated for an increase of $\delta_i = 1.65$ or 1.25 of their sampled speed parameter. The former was the increase in speed found in the empirical study by Zopluoglu, Kasli and Toton (2021); the latter was added to assess the effects of a milder increase.
5. The distributions were used to calculate the power functions in (18) with critical value z_{crit} set at the smallest value of z_{Pi} with right-tail probability $\Pr\{Z_{Pi} \geq z_{\text{crit}}|H_0\}$ not larger than level of significance $\alpha = .05$.
6. To account for the randomness in the IDs of the examinees with preknowledge, the procedure in Step 3 was replicated 1,000 times. The results were summarized plotting the 5th, 25th, 50th, 75th, and 95th percentiles in the distribution of the power functions.

The results in Figures 1–4 point at a statistical test that detects item compromise with nearly perfect power for a minimum of 10 out of the current 50 test takers with preknowledge of the more difficult or higher discriminating items. With more difficult but less discriminating items and the smaller increase in speed due to preknowledge, more test takers with preknowledge were necessary to detect the item as compromised. However, even then, some 20 test takers were sufficient to reach nearly perfect power. The presence of substantial correlation between speed and ability did not have any noticeable impact.

Relative to the results for the separate test with X_{Pi} in van der Linden (2022), the number of test takers with preknowledge required to obtain full power was reduced by some 50%. The power functions for the combined test were close to those for the separate RT-based test though, a fact taken to indicate that the RTs are the main driver of our new statistic Z_{Pi}. As revealed by absence of any substantial spread of the power functions in Figures 1–4 introduced by the random choice of test takers with preknowledge, the current test appears to have power functions that are extremely robust against variations of their regular speed and ability parameters, implying nearly

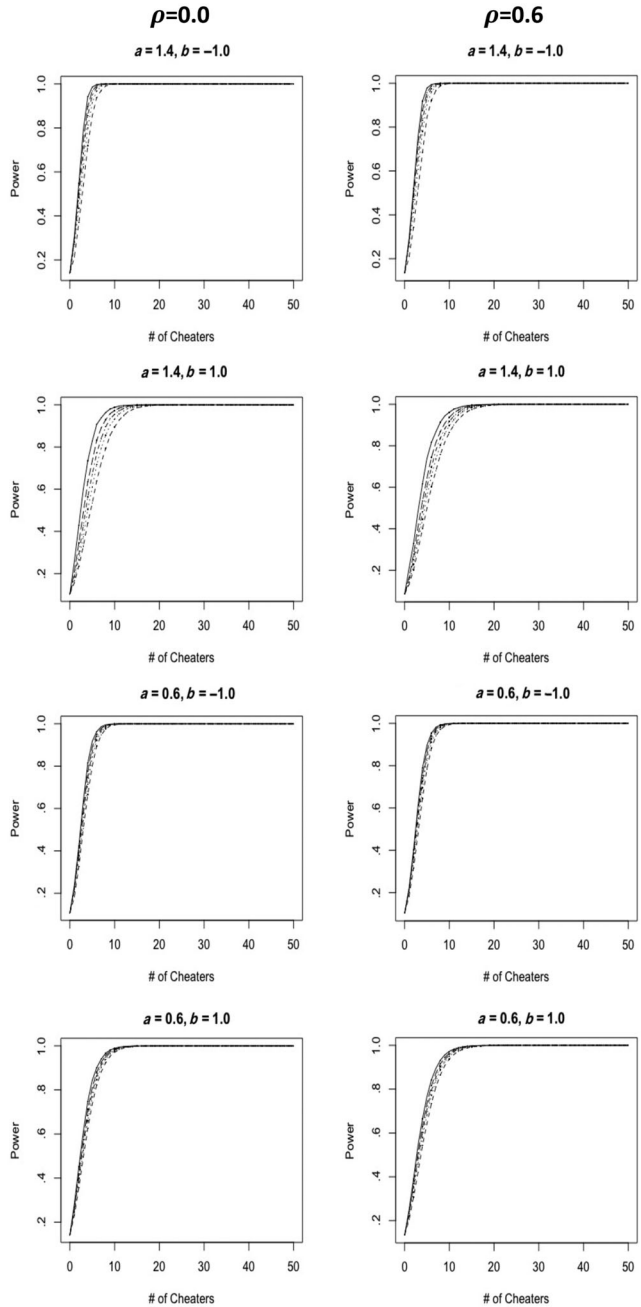

Figure 1. Quantiles of the distribution of power functions for the combined test with correlation between ability and speed equal to $\rho = .0$ and .6, discrimination parameter in the RT model equal to $\alpha_i = 2.3$, and an increase of speed due to preknowledge of $\delta_i = 1.65$. Curves from the right to the left are for the 5th, 25th, 50th, 75th, and 95th quantiles, respectively.

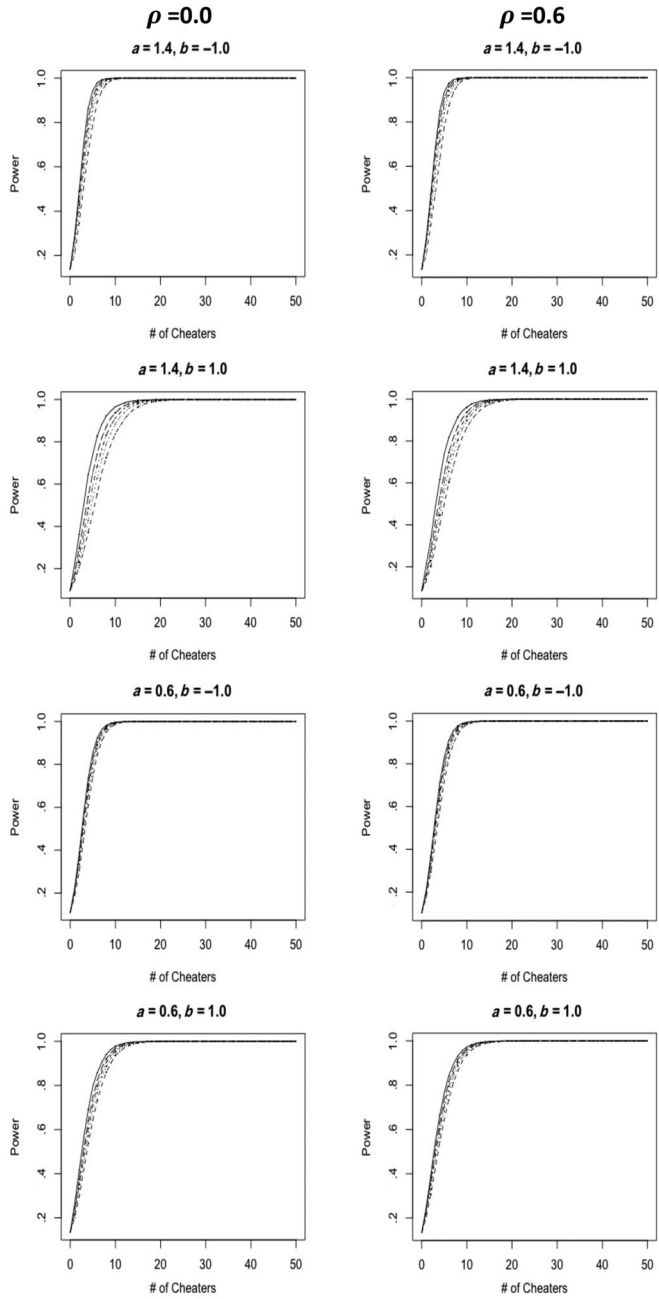

Figure 2. Quantiles of the distribution of power functions for the combined test with correlation between ability and speed equal to $\rho = .0$ and .6, discrimination parameter in the RT model equal to $\alpha_i = 2.3$, and an increase of speed due to preknowledge of $\delta_i = 1.25$. Curves from the right to the left are for the 5th, 25th, 50th, 75th, and 95th quantiles, respectively.

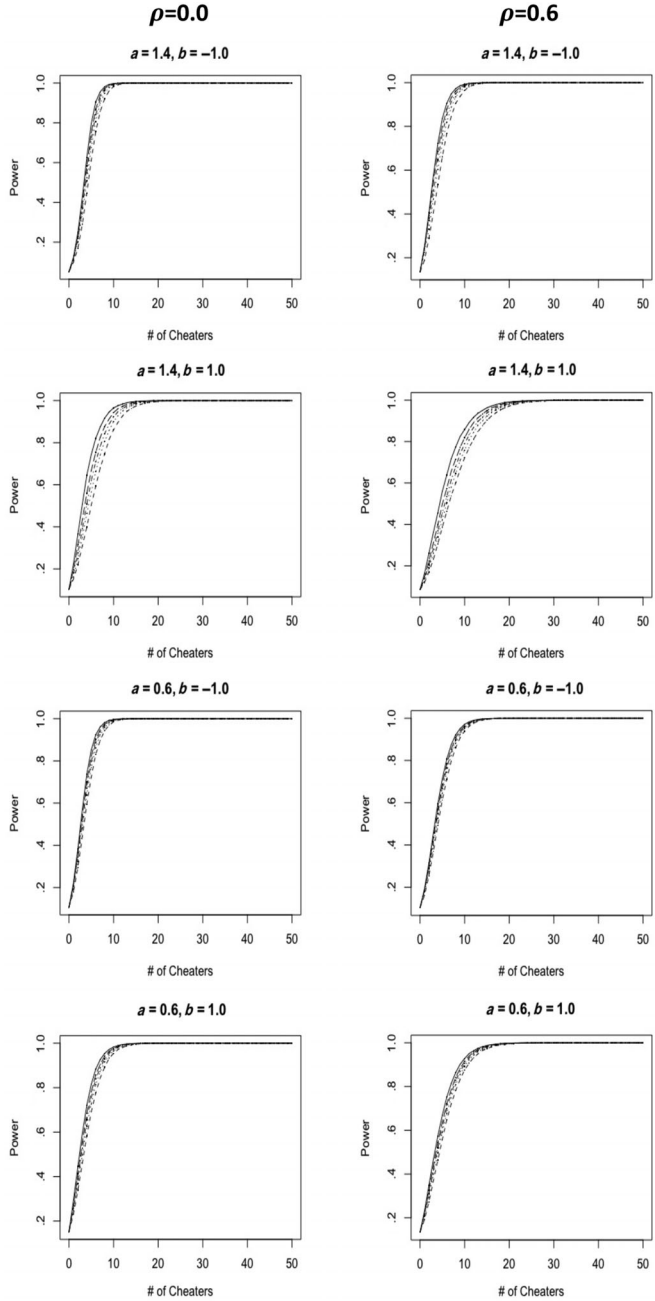

Figure 3. Quantiles of the distribution of power functions for the combined test with correlation between ability and speed equal to $\rho = .0$ and .6, discrimination parameter in the RT model equal to $\alpha_i = 1.4$, and an increase of speed due to preknowledge of $\delta_i = 1.65$. Curves from the right to the left are for the 5th, 25th, 50th, 75th, and 95th quantiles, respectively.

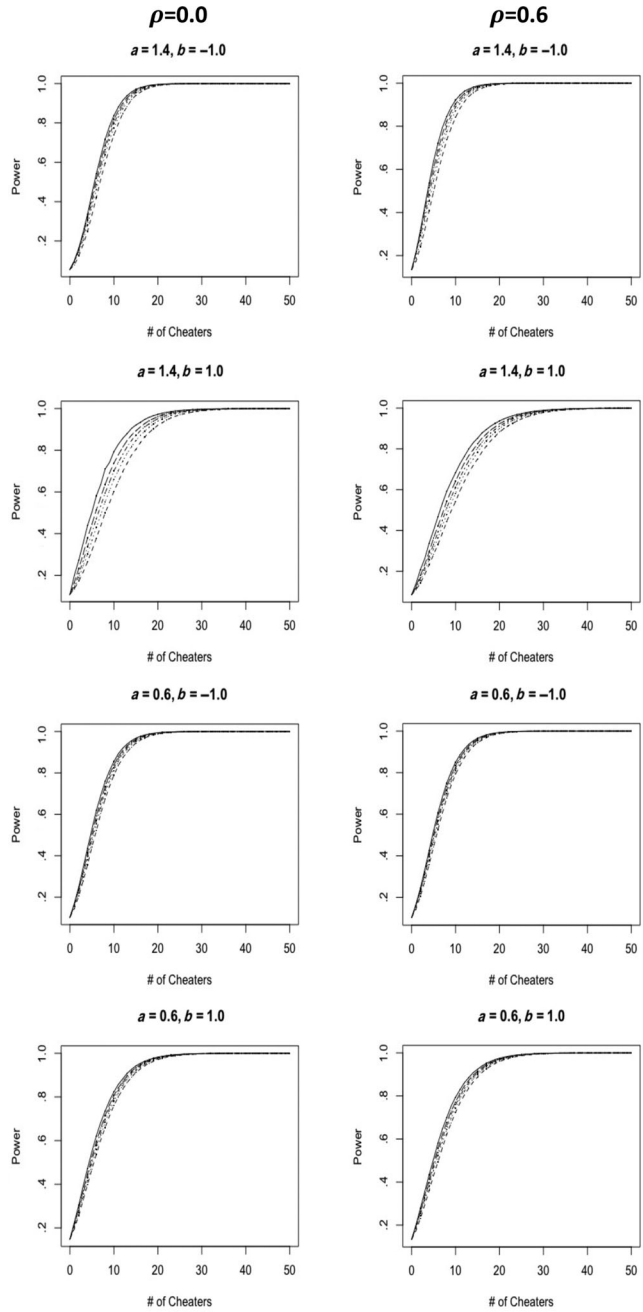

Figure 4. Quantiles of the distribution of power functions for the combined test with correlation between ability and speed equal to $\rho = .0$ and .6, discrimination parameter in the RT model equal to $\alpha_i = 1.4$, and an increase of speed due to preknowledge of $\delta_i = 1.25$. Curves from the right to the left are for the 5th, 25th, 50th, 75th, and 95th quantiles, respectively.

perfect certainty about the actual shape of the power functions for applications of the test. Also, the possible correlation between ability and speed parameters introduced in the current study did not have any noticeable impact on the shape of the power functions.

What About the Reverse Test?

An obvious remaining question is what would have been observed for a statistical test with the role of the responses and RTs reversed, that is, with the total time by the test takers on the item with responses flagged as suspicious as test statistic.

However, for dichotomous items, it is impossible to control significance level α for a version of the response-based test at the level of a single response. The only (nonrandomized) option left is to flag a correct response as significant and an incorrect one as not significant, using

$$Y_{Pi} = \sum_{p=1, U_{pi}=1}^{P} T_{pi}^{*} U_{pi}, \tag{20}$$

as test statistic; that is, the sum of the log-RTs by the test takers with a response $U_{pi} = 1$ on the item.

The statistic defines a family of mixture distributions of P normal variables with the RT distributions as components and the success probabilities as weights. The null distribution for a test of item compromise with Y_{Pi} as test statistic is the member of this family with the regular speed parameters τ_p for the components and regular success probabilities π_{pi} for each of the test takers, whereas the alternative distributions are members with inflated speed parameters and success probabilities for each of the $\gamma_i = 0, \ldots, P$ test takers with preknowledge, just as claimed in (A.2) and (A.7).

For a larger size of P, we could invoke a central limit theorem to claim a normal approximation to the family and proceed. However, further analysis reveals a fundamental problem for a test with (20) as statistic, which discourages its use.

Because of conditional independence of responses and RTs, the expected contribution to Y_{Pi} by each test taker is equal to

$$(\beta_i - \tau_{pi})\pi_{pi.} \tag{21}$$

Thus, assuming the test takers have operated independently during the test, the distribution of Y_{Pi} has a mean equal to

$$\sum_{p=1}^{P} (\beta_i - \tau_{pi})\pi_{pi}. \tag{22}$$

For the earlier test, the effect of an increase of τ_{pi} and π_{pi} as a result of item preknowledge was an increase in the probability of $Z_{pi} = 1$ and hence a positive shift in the distribution of Z_{Pi}. The separate effects of τ_{pi} and π_{pi} on Y_{Pi} are opposite though. If π_{pi} increases, (22) does increase, but an increase of τ_{pi} implies a decrease. As a consequence, dependent on the relative size of the two parameters, the net result of preknowledge could equally well be a shift of the distribution of Y_{Pi} to the left as the

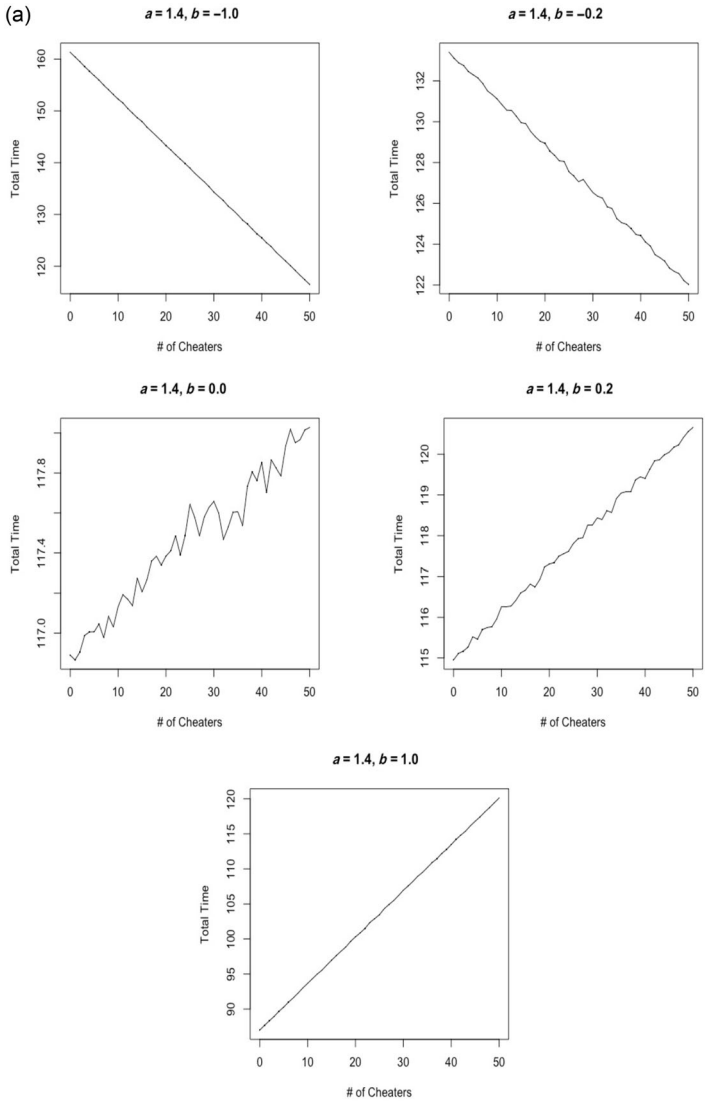

Figure 5. Examples of the mean of reverse statistic Y_{Pi} as a function of the number of cheaters γ_i for items with increasing values for their difficulty parameter b_i, discrimination parameter in the RT model equal to α=2.3, and an increase of speed due to preknowledge of $\delta_i = 1.65$.

right. A test based on the statistic would thus miss the MLR property necessary to decide if it should be left or right sided.

Figure 5 gives an example of how the mean of Y_{Pi} as a function of γ_i changes with an increase of the difficulty of the item. For a difficulty as low as $b_i = -1.0$, the mean decreases monotonically, just as necessary for a left-sided test. But for items with a difficulty approaching $b_i = 0$, the monotonicity of function disappears and

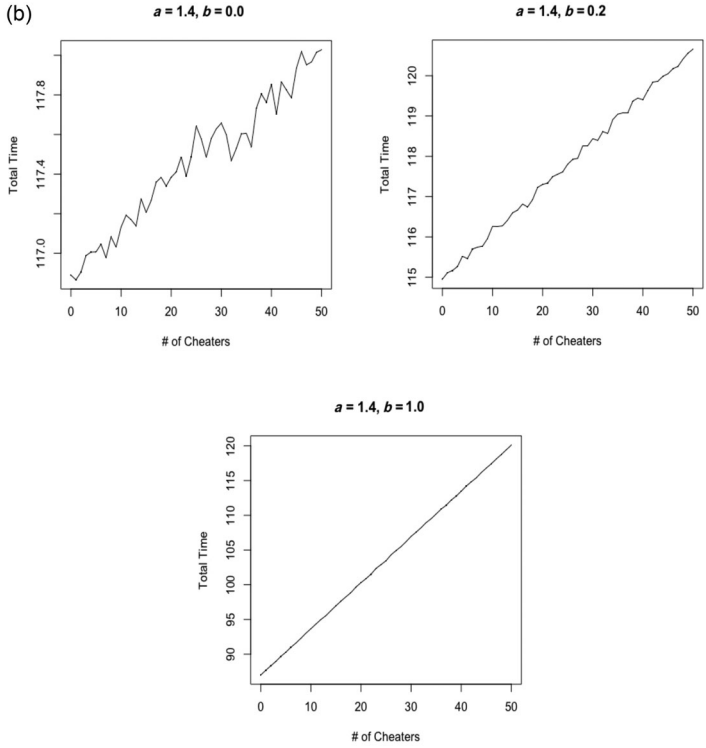

Figure 5. Continued

the function may even begin to increase. For the item with $b_i = 1$, the function is perfectly monotone again but now for an increase of the mean, just as one would want to see for a right-sided test.

Discussion

The proposed test assumes ability and speed parameters estimated from the regular items in the test. Though the estimation leads to loss of power, the actual loss is expected to be minor. For instance, as discussed in van der Linden (2022), for the choice of a window of $P = 50$ test takers and items with an average value for the discrimination parameter in the RT model of $\alpha_i = 1.85$ (middle between the items in Figure 1–4), the standard error in (8) is already as low as .01 for a speed parameter estimated from a test with 25 regular items, a result that holds no matter the time intensities β_i of the items. Estimation error in the ability parameter could be larger but, dependent on the size of the window, its impact is expected to average out across the test takers. In fact, the larger the number of test takers, the less the expected impact of estimation error on the test statistic. But, of course, a more detailed study is necessary to check the validity of these expectations.

The study should also address the choice of the level of significance for the test, as it is always possible to compensate for the impact of estimation error on the power

of the test by raising its significance level. In our examples, the level was chosen to be $\alpha = .05$ both for the test on the individual RTs of the test takers and the test based on the responses of all of them. In fact, as the price of replacing an item too early is expected to be much lower than the loss of reputation due to its item compromise, setting α higher than .05 might be the recommended choice for any testing program. Also, it is not necessary to choose the same level of significance for the RT-based and response-based parts of the test. The question of what the relative importance of the two levels on the power of the test is deserves research too.

The size of the window of test takers monitored is another important issue worth exploring. The choice of $P = 50$ in our examples should not be taken as a recommendation. Though the presence of some 10–20 test takers with preknowledge was sufficient to detect an item as compromised with nearly perfect power, there is no guarantee that in practice these numbers automatically show up in a window of this size. On the contrary, it is generally difficult to predict how fast test takers with preknowledge will arrive once an item is compromised. Their distribution over time depends, for instance, on the organization of the program, the intensity of the attempts to get hold of test items it is exposed to, and how well cheaters are organized—factors too specific to make any assumptions about. Our examples were intentionally planned to have 1,000 random replications of the positions of the test takers with preknowledge among the total of $P = 50$ test takers to assess the impact of the order in which they arrive. The minimal spread of the power functions in each of the examples in Figures 1–4 demonstrates that it is the mere number of test takers rather than the distribution of their arrival times that counts.

A test with high power is one thing, but a setup that promotes quick discovery of an item security breach is equally important to the defense of a testing program. One possible use of the proposed test may be to start the window at a time opportune to the program and check its items periodically for compromise while the window grows in size. The only data that need to be collected then are the probabilities π_{pi} and $F(t^*_{\text{crit}}; \tau_p, \alpha_i, \beta_i)$ for each test taker that arrives. The test itself requires only a launch of the Lord-Wingersky algorithm. Other setups are possible, so more detailed study is required here as well.

A few final warnings are necessary. First, with cheating on tests nowadays becoming more of an organized (and even industrialized) operation, test takers with preknowledge of items may be coached to pace themselves in order to avoid being detected by forensic screening. The general effect of this coaching is a lower increase in speed than the empirical values used in our examples and larger samples required to achieve the same power. Additional analyses are necessary to catch this new type of cheaters. One starting point is to check on irregularities in the RTs rather than their decrease. It is generally difficult for test takers who have already familiarized themselves with items to mimic the behavior of naive test takers who see them for the first time, particularly because, as found by the current authors in earlier studies with the RT model in (3), the time intensities of items in a test easily differ by a factor of 5–10. Second, as one of our reviewers indicated, items may be disclosed with incorrect answers. If this happens, a success probability of $\pi_{pi} = 1$ is too high and the actual power of the test is lower than suggested by the results from our simulation study. We wonder, however, how frequent this may happen. It is only for high-stakes

tests that test takers are willing to pay for disclosed items. And in order to prepare for such tests, we expect them to practice the items, or, as just discussed, even attend coaching sessions. It is important to note that the choice of probability π_{pi} only has an impact under the alternative hypothesis; critical value z_{crit} in (15) is independent of it. Consequently, the only option for a testing program to deal with the presence of disclosed items with incorrect answers is to increase the level of significance of the test, something that, as just indicated, may be recommendable anyhow. Third, it is important to remember that a significant statistical test is not a sufficient condition for item compromise to be true. Other reasons for unexpected correct responses or higher speed may exist. For example, a higher actual speed on an item than expected could also be the result of guessing if it tends to be located toward the end of a speeded test. Likewise, a higher probability of success than expected could be the result of a test taker just working at a lower speed on an occasional item, a result we should particularly be aware of when the item tends to be located toward the beginning of the test. It is therefore important to supplement positive results on any statistical test of item compromise with reports of observed irregular behavior by proctors, visits to websites offering stolen items prior to the test, or communication detected between test takers with suspected preknowledge of the same set of items.

Acknowledgments

The first author received funding for this study from the Law Schools Admission Council (LSAC). The opinions and conclusions contained in this article do not necessarily reflect the policy and position of the LSAC.

References

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. (2010). Automatic detection of answer copying via Kullback-Leibler convergence and K-index. *Applied Psychological Measurement*, *34*, 379–392.

Belov, D. I., & Cubbellotti, S. (2017). *Detecting examinees with aberrant answer changes in CBT via posterior shift*. Presented at the Conference on Test Security, Madison, WI.

Belov, D. I., & Toton, S. L. (2022). Detecting examinees with item preknowledge on real data. *Applied Psychological Measurement*, *41*(5), 338–352. doi:10.1177/01466216221084202

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.). Pacific Grove, CA: Duxbury.

Chen, Y., Lee, Y. H., & Li, X. (2021). Item pool quality control in educational testing: Change point model, compound risk, and sequential detection. *Journal of Educational and Behavioral Statistics*, *47*(3). doi: 10.3102/107699862110590085

Choe, E. M., & Chang, H. H. (2014, April). *Utilizing response time in sequential detection of compromised items*. Annual Meeting of the National Council on Measurement in Education, Philadelphia, PA.

Choe, E. M., Zhang, J., & Chang, H. H. (2018). Sequential detection of compromised items using response times in computerized adaptive testing. *Psychometrika*, *83*, 650–673.

Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education*, *9*, 47–64.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. New York: Springer.

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).

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.

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

Sinharay, S. (2017b). Which statistic should be used to detect item preknowledge when the set of compromised items is known? *Applied Psychological Measurement*, *41*, 403–421.

Sinharay, S. (2020). Detection of item preknowledge using response times. *Applied Psychological Measurement*, *44*, 376–392.

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.

Toton, S. L., & Maynes, D. D. (2019). Detecting examinees with pre-knowledge in experimental data using conditional scaling of response times. *Frontiers in Education*, *4*, 1–18.

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

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

van der Linden, W. J. (2018). Item and test security. In W. J.van der Linden (Ed.), *Handbook of item response theory: Volume 3. Applications* (pp. 267–293). Boca Raton, FL: Chapman & Hall/CRC.

van der Linden, W. J. (2022). Two statistical tests for the detection of item compromise. *Journal of Educational and Behavioral Statistics*, *47*(4), 485–504. https://doi.org/10.3102/10769986221094789

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

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, X., & Liu, Y. (2020). Detecting compromised items using information from secure items. *Journal of Educational and Behavioral Measurement*, *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.
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.

Appendix A: Two Separate Tests of Item Compromise

For the response-based test, the null and alternative hypotheses in (9) and (10) are further specified as

$$H_0 : \Pr\{U_{pi} = 1\} = \pi_{pi} \text{ for all test takers}, \tag{A.1}$$

and

$$H_1 : \Pr\{U_{pi} = 1\} > \pi_{pi} \text{ for } \gamma_i \text{ of the test takers}, \tag{A.2}$$

where π_{pi} is the probability of success in the response model in (1)–(2) for the test takers' regular ability parameters demonstrated on the uncompromised items in the test.

As the test takers are assumed to work independently during testing, the probability of observing a response vector $\boldsymbol{u_i} \equiv (u_{1i}, \ldots, u_{Pi})$ follows from (1) as

$$\Pr\{\boldsymbol{U}_i = \boldsymbol{u}_i; P, \boldsymbol{\pi}_i\} = \prod_{p=1}^{P} \pi_{pi}^{u_{pi}} (1 - \pi_{pi})^{1-u_{pi}}, \tag{A.3}$$

where $\boldsymbol{\pi}_i \equiv (\pi_{1i}, \ldots, \pi_{Pi})$. The probability mass function (pmf) of the total number of correct responses X_{Pi} by the P test takers on the item is equal to

$$f(x; P, \boldsymbol{\pi}_i) = \Pr\{X_{Pi} = x; P, \boldsymbol{\pi}_i\}$$
$$= \begin{cases} \sum_{\sum u_{pi} = x} \prod_{p=1}^{P} \pi_{pi}^{u_{pi}} (1 - \pi_{pi})^{1-u_{pi}}, & x = 0, 1, \ldots, P; \\ 0, & \text{otherwise.} \end{cases} \tag{A.4}$$

Distributions with this type of pmf belong to the compound binomial family, with the null and alternative distributions as members with different parameter values. For the null hypothesis, $\boldsymbol{\pi}_i$ is the vector with the probabilities of success for the test takers' regular ability parameters. The alternative hypothesis claims $\gamma_i > 0$ of these probabilities to be for test takers with preknowledge of the item. If these test takers can be assumed to know the correct answers to the items as well, an obvious choice for them is $\pi_{pi} = 1$.

The family has the important property of a monotone likelihood ratio (MLR) for the vector of response probabilities in the number of correct responses (Romero, Ri-

ascos & Jara, 2015). Consequently, the same holds for likelihood ratio $f(x_{Pi}; \gamma_i > 0)/f(x_{Pi}; \gamma_i = 0)$, which is a monotonically increasing function of x_{Pi}. The property implies a right-sided level α test that rejects $H_0 : \gamma_i = 0$ in favor of $H_1 : \gamma_i > 0$ when

$$X_{Pi} \geq x_{\text{crit}}, \tag{A.5}$$

where x_{crit} is the smallest value of x_{Pi} with right-tail probability $\Pr\{X_{Pi} \geq x_{\text{crit}}|H_0\}$ not larger than level of significance α.

Due to their combinatorial complexity, it may seem difficult to compute the probabilities in (A.4). But they are easily calculated using the recursive algorithm in Lord and Wingersky (1984) widely used in the test-theory literature to calculate the probabilities of number-correct score distributions.

For the RT-based test, **the two basic hypotheses in (9)–(10) specialize to

$$H_0 : \tau_{pi} = \tau_p \text{ for all P test takers;} \tag{A.6}$$

$$H_1 : \tau_{pi} > \tau_p \text{ for } \gamma_i \text{ of the test takers,} \tag{A.7}$$

where τ_p is the regular speed parameter for the test takers as demonstrated on the uncompromised items in the test.

As the test takers are assumed to have worked independently during the test and the sum of normal random variables is normal as well, it follows from (4)–(6) that the probability density function (pdf) of the sum of the log-RTs, T^*_{Pi}, by the P test takers belongs to the normal family with mean

$$\mu_{Pi} = P\beta_i - \sum_{p=1}^{P} \tau_{pi} \tag{A.8}$$

and variance

$$\sigma^2_{Pi} = P\alpha_i^{-2}. \tag{A.9}$$

As τ_{pi} is the only unknown parameter, the family has a known variance but unknown mean. The null distribution is the member with the regular speed parameters $\tau_{pi} = \tau_p$ for all P examinees as demonstrated on the other items in the test. The alternative distributions have γ_i of their speed parameters equal to $\tau_{pi} + \delta_i > \tau_p$, where δ_i is the increase in speed due to knowledge of the item. A recent empirical study by Zopluoglu, Kasli, and Toton (2021) reported an increase of $\delta_i = 1.65$ for test takers on items to which they had access prior to the test.

The family is known to have the same property of MLR for its mean (Casella & Berger, 2002, example 8.3.15). As (A.8) decreases with an increase of τ_{pi} for each of the test takers, it follows that $f(t^*; \gamma_i > 0)/f(t^*; \gamma_i = 0)$ is a monotonically decreasing function of t^*. The property implies a left-sided test of H_0 against H_1. Thus, H_0 should be rejected in favor of H_1 when

$$T^*_{Pi} < t^*_{\text{crit}}, \tag{A.10}$$

where

$$t^*_{\text{crit}} = P\beta_i - \sum_{p=1}^{P} \tau_p + z_a P^{1/2} \alpha_i^{-1}, \tag{A.11}$$

and z_α is the αth quantile of the standard normal distribution.

According to the Neyman-Pearson lemma, both tests are uniformly most powerful (UMP) for known ability and speed parameters (Casella & Berger, 2002, theorem 8.3.12), a condition that holds only approximately in practice, of course. For empirical examples of the power functions for the two test and a discussion of the impact of the estimation of the examinees parameters, see van der Linden (2022).

Authors

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 primary research interests include test theory, applied statistics, and research methods

DMITRY I. BELOV is a Principal Computer Scientist, Department of Assessment Sciences, Law School Admission Council, 662 Penn Street, Newtown, PA 18940; dbelov@lsac.org. His primary research interests include statistical methods for detecting cheating on tests and item difficulty modeling.