A Bivariate Lognormal Response-Time Model for the Detection of Collusion Between Test Takers

Wim J. van der Linden
CTB/McGraw-Hill

A bivariate lognormal model for the distribution of the response times on a test by a pair of test takers is presented. As the model has parameters for the item effects on the response times, its correlation parameter automatically corrects for the spuriousness in the observed correlation between the response times of different test takers because of variation in the time intensities of the items. This feature suggests using the model in a routine check of response-time patterns for possible collusion between test takers using an estimate of the correlation parameter or a statistical test of a hypothesis about it. Closed-form expressions for the maximum-likelihood estimations of the model parameters and a Lagrange multiplier test for the correlation parameter are presented. As in any type of statistical decision making, results from such procedures should be corroborated by evidence from other sources, for example, results from a response-based analysis or observations during the test session. The effectiveness of the model in removing the spuriousness from correlated response times is illustrated using empirical response-time data.

Keywords: cheating; Lagrange multiplier test; lognormal model; maximum-likelihood estimation; response time

One of the increasing concerns of producers and consumers of test scores is fraudulent behavior during the test. Such behavior is more likely to be observed when the test takers’ stakes are high, such as in admission, licensing, and certification testing. A simple form of cheating is one test taker attempting to copy the answers from another. This may happen without the source noticing. The classical form of this type of cheating is one test taker peeking at another’s answer sheet. A more modern form arises when computerized tests are administered in a regular classroom setting. Computer screens now have such a high resolution that, unless special arrangements are made, students can easily read the screens of others a few rows in front of them.

This study received funding from the Law School Admissions Council (LSAC). The opinions and conclusions contained in this article are those of the author and do not necessarily reflect the policy and position of LSAC.
Test takers may also plan to collude during the test. A classical form of collusion is the signaling of question and answer numbers between test takers using a prearranged code (e.g., number of silent finger taps). More advanced communication is possible in the form of electronic transmission of data. One obvious instance is communication between test takers through a local network or over the internet when their PCs have not been locked down properly. However, more sophisticated forms have now become possible as a result of the rapid miniaturization of electronic devices. For instance, some digital pens now look like regular pens and automatically record handwritten responses or notes. The author would not be surprised if soon test takers would be caught transmitting such information wirelessly to others.

The traditional weapon of the testing industry against collusion is analysis of the agreement between the response patterns of the suspected test takers. The key quantity in all current statistics for detecting answer copying on multiple-choice items—traditionally one of the item formats most vulnerable to cheating—is the number of matching alternatives between test takers (Angoff, 1974; Frary, Tideman, & Watts, 1977; Holland, 1996; Lewis & Thayer, 1998; Sotaridona & Meijer, 2002, 2003; Sotaridona, van der Linden, & Meijer, 2006; van der Linden & Sotaridona, 2004, 2006; Wollack, 1997; Wollack & Cohen, 1998). However, these statistics differ in more or less subtle aspects. For example, some of them ignore matching correct alternatives, the idea being that a string of them may also be indicative of high-ability test takers who worked independently, whereas a string of matching incorrect alternatives is much less likely to occur under normal circumstances. Another difference resides in the null distributions postulated for these statistics. Some of them make an appeal to large-sample normality after some form of standardization, whereas others model the process of answer copying and derive the test statistic from the model.

It is not our intention to review these tests of answer copying or advocate a position with respect to the different assumptions underlying them. Readers interested in our view should refer to the introductory section of van der Linden and Sotaridona (2006). Instead, the goal of this research was to find an alternative method of detecting collusion between test takers based on their response times (RTs). RTs are a valuable extra source of information about the behavior of test takers during the test. For example, they have the advantage of being continuous rather than dichotomous and are therefore more informative about the size of possible aberrances than responses.

An even more attractive feature of RTs is that their use is not restricted to any specific response format. Particularly, it does not suppose the multiple-choice format for which the statistical tests above were developed. In fact, the more complicated the response format, the more informative the RTs. This point can be illustrated by the constructed-response format, for which we could not only analyze the total RT but also the time elapsed, for example, until a specific concept was introduced or a necessary tool was used. Of course, it will be more
difficult to automate RT analyses, the closer we get to the level of the full set of
time stamps in the test takers’ logfiles. However, the possibilities are seemingly
endless.

Finally, because the focus is no longer on responses, we avoid the statistical
subtleties involved in the decision on what information the statistical test should
condition: the number of incorrect alternatives by the source, his or her entire pat-
tern of incorrect alternatives, or even the two full patterns of responses for the
pair of test takers (Lewis, 2006). Such decisions become especially difficult
when it is impossible to identify one test taker as the source and the other as the
copier.

However, RT-based procedures for detecting collusion may seem to have dis-
advantages. Clearly, they can only be applied to tests administered in a comput-
erized mode. However, we believe this to be a short-term disadvantage.
Computerized testing offers numerous improvements over paper-and-pencil test-
ing, such as the use of innovative item types, immediate scoring, and adaptive
item selection. We therefore expect it soon to be the standard mode of testing.
Because collusion is only possible when items between test takers are the same,
this type of cheating is not a serious danger to adaptive testing, though. For RT-
based procedures to detect aberrances more typical of adaptive testing, such as
memorization and preknowledge of some of the items, see van der Linden and
Guo (2008).

A more serious disadvantage may be test takers becoming aware of the fact
that their RTs are checked and trying to fake realistic RTs to hide their collusion.
However, we expect this to be too difficult for them, especially because they
have to do this in real time. The expected RTs on the items in a typical opera-
tional test easily differ by a factor 5–8. For test takers trying to hide their collu-
sion, it would be quite a challenge to find out what their patterns should be on a
set of test items they have not seen before while keeping an eye on the clock, still
solving the items, and avoiding running out of time. A more effective strategy for
them might be to plan ahead. One possibility that has come to the author’s mind
is an odd-even strategy, in which one test taker works on the odd and the other on
the even items and they periodically exchange the answers. Such strategies may
not be detected by the statistics presented in this manuscript but would always
leave a trail of RTs that could be picked up by supplemental analyses. In fact,
this particular trail would immediately be flagged by the procedure for the detec-
tion of aberrances in individual RT patterns in van der Linden and Guo (2008).

As any other area of statistical decision making, detection of collusion
between test takers is prone to Type 1 and Type 2 errors—here: the errors of false
accusations of cheating and incorrect assumptions of regular behavior. The for-
mer hurt innocent test takers; the latter are unfair to fellow test takers, especially
if they are competing with the cheaters for scarce positions. In addition, cheating
always undermines the credibility of the testing program.
The question of how to balance these errors is a policy issue that transcends the statistical aspects of the problem. But whatever the policy, it is always beneficial to take steps to reduce the likelihood of both types of errors to the maximum extent possible. One important step is to conduct response-based and RT-based procedures independently, and then check whether their conclusions agree. In addition, the decisions should be corroborated by additional observations (e.g., reports by proctors and analyses of logfiles). In the absence of such evidence, a practical approach already followed by some test organizations is to suspend the decision and invite the suspects to take the test again under controlled conditions.

1. RT Models

The analog of a string of matching responses between two test takers is a string of correlating RTs. The detection of such strings should be based on a statistical model for RT distributions under regular test behavior; otherwise, we might easily make wrong inferences. For example, as already indicated, RT distributions on test items depend heavily on their labor or time intensity. Because of this, the RTs of different test takers on the same set of test items always show a tendency to correlate, simply because of the variation in the time intensity from one item to the next. Likewise, test takers vary in the speed at which they operate. Ignoring this factor might lead to the confusion of speed with collusion as the cause of similar RTs. For empirical examples of spurious relations in patterns of observed RTs, see van der Linden (in press) and van der Linden and Guo (2008).

A model that does take item and person effects into account is the lognormal RT model (van der Linden, 2006). The core of the model is an equation that defines the speed of labor by a test taker on an item. Let $B_i^*$ denote the (unknown) amount of labor required to solve item $i$ and $t_{ij}$ the RT by test taker $j$ on this item. Observe that this RT is the time elapsed during labor. Hence, the (average) speed of labor $\tau_j^*$ by test taker $j$ on item $i$ is

$$\tau_j^* = \frac{B_i^*}{t_{ij}}.$$  \hspace{1cm} (1)

This definition follows the format of any definition of speed as the rate of change of some substantive measure with respect to time. For example, the well-known definition of the speed of motion of a body from physics has the same form as Equation 1 but with the amount of labor in the numerator replaced by the distance traveled. For more on this fundamental equation of RT modeling, see van der Linden (in press).

Observe that Equation 1 is equivalent to

$$t_{ij} = \frac{B_i^*}{\tau_j^*},$$  \hspace{1cm} (2)
which shows that the equation actually decomposes one (known) RT into two (unknown) effects for the person and the item. The lognormal model follows from Equation 2 in two obvious steps: (a) a logarithmic transformation of the RTs to remove the typical skewness from their distributions and (b) the addition of an item-specific random term to allow for the randomness of RTs. The result is

$$\ln T_{ij} = \beta_i - \tau_j + \varepsilon_i, \varepsilon_i \sim N(0, \sigma_i^2),$$

(3)

where $\beta_i$ and $\tau_j$ are now parameters for the item and person effects on a logarithmic scale and $\alpha_i$ can be interpreted as an item discrimination parameter. Equivalently, $T_{ij}$ follows a lognormal distribution with density,

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

(4)

where $\tau_j \in \mathbb{R}$ is the speed at which test taker $j$ operates on the test, $\beta_i \in \mathbb{R}$ is the time or labor intensity of item $i$, and $\alpha_i \in \mathbb{R}^+$ is its discrimination parameter. Factor $1/t_{ij}$ in Equation 4 is because of the change of variable in the density. The model is not yet identifiable but a convenient way to obtain identifiability is to set

$$\mu_\tau = 0,$$

(5)

for the population of test takers on which the items are calibrated.

For statistical issues related to parameter estimation and model validation, which can be performed efficiently by embedding the model in a hierarchical framework along with a regular response model, see Fox, Klein Entink, and van der Linden (2007); Klein Entink, Fox, and van der Linden (2009); and van der Linden (2006, 2007). A review of empirical applications of the model is given in van der Linden (2009).

### 1.1 Bivariate Lognormal Model

To analyze the RTs of pairs of test takers for possible collusion when their regular behavior can be assumed to follow the lognormal model in Equation 4, we need an extension of the model for the joint distribution of their RTs on a fixed item. Obviously, for a pair of test takers $(j, k)$, the bivariate generalization of Equation 3 for the distribution of $(\ln T_{ij}, \ln T_{ik})$ on item $i$ has density

$$f(\ln t_{ij}, \ln t_{ik}; \tau_j, \alpha_i, \beta_i, \rho_{jk}) = \frac{\alpha_i^2}{2\pi\sqrt{1 - \rho^2_{jk}}} \exp\left\{ -\frac{1}{2(1 - \rho^2_{jk})} \left[ \psi_{jk}^2 - 2\rho_{jk}\psi_i\psi_{ik} + \psi_{ik}^2 \right] \right\},$$

(6)

where

$$\psi_{ij} = \alpha_i[\ln t_{ij} - (\beta_i - \tau_j)].$$

(7)
Equivalently \((T_{ij}, T_{ik})\) has a bivariate lognormal density

\[
f(t_{ij}, t_{ik}; \tau_j, \tau_k, \alpha_i, \beta_i, \rho_{jk}) = \frac{\alpha^2}{t_{ij}t_{ik}2\pi\sqrt{1 - \rho^2_{jk}}} \exp\left\{ \frac{-1}{2(1 - \rho^2_{jk})} (\psi^2_{ij} - 2\rho_{jk}\psi_{ij}\psi_{ik} + \psi^2_{ik}) \right\}.
\]  

(8)

Factor \(1/t_{ij}\) in Equation 4 generalizes to \(1/t_{ij}t_{ik}\) in Equation 8 as a result of the two separate log transformations of \(T_{ij}\) and \(T_{ik}\). For \(\rho_{jk} = 0\), Equations 6 and 8 degenerate to the product of the two densities for the normal and lognormal representations of the univariate model in Equations 3 and 4, respectively. For computational convenience, we will capitalize on the representation in Equation 6.

Note that \(\rho_{jk}\) is a parameter for the degree to which the RTs by two fixed test takers agree. It is thus independent of the RTs of any of the other test takers. This fact is important because several of the statistical tests for matching alternatives on multiple-choice tests referred to earlier are population dependent and lead to different decisions when the same pair of test takers would be included in different populations. In addition, note that \(\rho_{jk}\) is the correlation between \(\psi_{ij}\) and \(\psi_{ik}\) in Equation 7. Because the impact of the identifiability restriction in Equation 5 on \(\beta_i\) and \(\tau_j\) on \(\psi_{ij}\) cancels, the use of the population mean in Equation 5 does not make any of the inferences about \(\rho_{jk}\) population dependent.

Another important feature of the model is its automatic correction of \(\rho_{jk}\) for the time intensities \(\beta_i\) of the items. As a result, \(\rho_{jk}\) is insensitive to spuriousness in the observed correlation between the logtimes of two test takers because of the variation in the time intensity between the items. The model also allows us to compare the speed \(\tau_j\) of different test takers. This is important because one would expect test takers who collude to work at approximately at the same speed. Although there may be several other reasons why test takers have similar speed, this expectation makes \(\tau = (\tau_j, \tau_k)\) a second potential parameter of interest. We will return to this point later.

2. Estimation of \(\rho_{jk}\)

It is assumed that the items have been calibrated with enough precision before their operational use and that we can therefore treat their parameters as known. Thus, for test takers \(j\) and \(k\) on items \(i = 1, \ldots, n\), the only unknown parameters are the parameters of interest, \(\rho_{jk}\) and \(\tau = (\tau_j, \tau_k)\). In an abuse of notation, let \(\ln t_j = (\ln t_{1j}, \ldots, \ln t_{nj})\) and \(\ln t_k = (\ln t_{1k}, \ldots, \ln t_{nk})\). Hence, the log likelihood of \((\rho_{jk}, \tau)\) can be written as follows:

\[
\ell(\rho_{jk}, \tau; \ln t_j, \ln t_k) = \text{const} - \frac{n}{2} \ln(1 - \rho^2_{jk}) - \sum_{i=1}^{n} \frac{1}{2(1 - \rho^2_{jk})} (\psi^2_{ij} - 2\rho_{jk}\psi_{ij}\psi_{ik} + \psi^2_{ik}).
\]  

(9)
From Equation 9, using the first-order derivatives in Equations A1, A2, and A3 in the appendix, the likelihood equations for \((\rho_{jk}, \tau)\) follow as

\[
\frac{\partial \ell(\rho_{jk}, \tau; \ln t_j, \ln t_k)}{\partial \tau_j} = \frac{1}{1 - \rho_{jk}^2} \sum_{i=1}^{n} \gamma_i (\rho_{jk} \Psi_{ik} - \Psi_{ij}) = 0, \tag{10}
\]

\[
\frac{\partial \ell(\rho_{jk}, \tau; \ln t_j, \ln t_k)}{\partial \tau_k} = \frac{1}{1 - \rho_{jk}^2} \sum_{i=1}^{n} \gamma_i (\rho_{jk} \Psi_{ij} - \Psi_{ik}) = 0, \tag{11}
\]

\[
\frac{\partial \ell(\rho_{jk}, \tau; \ln t_j, \ln t_k)}{\partial \rho_{jk}} = \frac{n \rho_{jk}}{1 - \rho_{jk}^2} - \sum_{i=1}^{n} \frac{\rho_{jk} \Psi_{ij}^2 - \rho_{jk}^2 \Psi_{ij} \Psi_{ik} + \rho_{jk} \Psi_{ik}^2 - \Psi_{ij} \Psi_{ik}}{(1 - \rho_{jk}^2)^2} = 0. \tag{12}
\]

The first two equations yield

\[
\rho_{jk} \sum_{i=1}^{n} \gamma_i^2 [\ln t_{ik} - (\beta_i - \tau_k)] = \sum_{i=1}^{n} \gamma_i^2 [\ln t_{ij} - (\beta_i - \tau_j)], \tag{13}
\]

\[
\rho_{jk} \sum_{i=1}^{n} \gamma_i^2 [\ln t_{ij} - (\beta_i - \tau_j)] = \sum_{i=1}^{n} \gamma_i^2 [\ln t_{ik} - (\beta_i - \tau_k)]. \tag{14}
\]

Note the symmetry between Equations 13 and 14. Hence, they only hold when the two sums are equal to 0 or \(|\rho_{jk}| = 1\). The first condition gives

\[
\hat{\tau}_j = \frac{\sum_{i=1}^{n} \gamma_i^2 (\beta_i - \ln t_{ij})}{\sum_{i=1}^{n} \gamma_i^2}, \tag{15}
\]

\[
\hat{\tau}_k = \frac{\sum_{i=1}^{n} \gamma_i^2 (\beta_i - \ln t_{ik})}{\sum_{i=1}^{n} \gamma_i^2}. \tag{16}
\]

The second condition can be ignored because the interest is in a nondegenerate bivariate distribution. In addition, \(\rho_{jk} = 1\) involves identical RTs and in this degenerate case, Equations 15 and 16 already give identical results.

Following an argument in Lehmann (1999, ex. 7.5.4), the solution of Equation 12 is

\[
\hat{\rho}_{jk} = \frac{\sum_{i=1}^{n} \hat{\Psi}_{ji} \hat{\Psi}_{ik}}{\left(\sum_{i=1}^{n} \hat{\Psi}_{ji}^2 \sum_{i=1}^{n} \hat{\Psi}_{ik}^2\right)^{1/2}}, \tag{17}
\]

where \(\hat{\Psi}_{ji}\) and \(\hat{\Psi}_{ik}\) are the estimates in Equation 6 with Equations 15 and 16 substituted for the \(\tau\)s.
The maximum-likelihood estimator (MLE) of $\tau_j$ weighs the observations $\beta_i - \ln t_{ij}$ by the square of the item discrimination parameters $\alpha_i$. Its expression thus has the well-known form of a precision-weighted average. The averaged quantity is $\beta_i - \ln t_{ij}$, which is the negative of the logtimes corrected for the time intensities of the items (the negative is due to the fact that $\tau_j$ is a speed and not a slowness parameter). The MLE of $\rho_{jk}$ appears to have the regular form of a sample covariance over the product of its two standard deviations. However, these quantities are not calculated over the logtimes but their estimated residuals in Equation 7.

3. Test of Hypothesis on $\rho_{jk}$

In a routine screening of test data, we could not only calculate the MLE of $\rho_{jk}$ but also conduct a formal statistical test of a hypothesis about the parameter. The main difference between the two procedures is that the latter allows for the uncertainty in the parameter estimates.

We discuss the test of

$$H_0 : \rho_{jk} = 0$$

against

$$H_1 : \rho_{jk} > 0.$$  \hspace{1cm} (18)

This test has the advantage of a statistic that reduces to a simple expression and allows us to directly follow the argument on which it is based. The generalization to a test of the more general hypothesis $H_0 : \rho_{jk} = c$ against $H_1 : \rho_{jk} > c$ is outlined in the appendix. The generalization is straightforward and allows us to check whether $\rho_{jk}$ differs substantially enough from zero to warrant concerns about possible collusion between the test takers, for instance, in situations when the test of Equation 18 against Equation 19 appears to have too much power and may detect negligible differences from $\rho_{jk} = 0$. For such cases, see below.

The proposed test is a Lagrange multiplier (LM) test. For a general introduction to this type of test and a discussion of their favorable statistical properties, the reader is referred to Lehmann (1999, sect. 7.7) or Silvey (1975, sect. 7.4). The use of LM tests for the diagnosis of violations of various IRT models has been popularized by recent work by Glas and his associates (e.g., Glas, 1999; Glas & Dagohoy, 2007; Glas & Suárez Falcón, 2003; van der Linden & Glas, in press). A practical feature of LM tests is that their statistics have to be evaluated only at the MLEs of the parameters in the null model and more complicated computation of any of the estimates of the alternative parameters is avoided.

The statistic for the LM test of Equation 18 against Equation 19 with unknown parameters $\tau = (\tau_j, \tau_k)$ can be written as follows:
\[
\text{LM}(p_{jk}) = \frac{h(p_{jk})^2}{h(p_{jk}, p_{jk}) - H(\tau, p_{jk})' H(\tau, \tau)^{-1} H(\tau, p_{jk})}_{\tau = \hat{\tau}, p_{jk} = 0},
\]

where

\[
h(p_{jk}) = \frac{\partial \ell(p_{jk}, \tau; t_j, t_k)}{\partial p_{jk}},
\]

\[
h(p_{jk}, p_{jk}) = -\frac{\partial^2 \ell(p_{jk}, \tau; t_j, t_k)}{\partial p_{jk}^2},
\]

\[
H(\tau, \tau) = \begin{pmatrix}
\frac{\partial^2 \ell(p_{jk}, \tau; t_j, t_k)}{\partial \tau_j^2} & \frac{\partial^2 \ell(p_{jk}, \tau; t_j, t_k)}{\partial \tau_j \partial \tau_k} \\
\frac{\partial^2 \ell(p_{jk}, \tau; t_j, t_k)}{\partial \tau_k \partial \tau_j} & \frac{\partial^2 \ell(p_{jk}, \tau; t_j, t_k)}{\partial \tau_k^2}
\end{pmatrix},
\]

and

\[
H(\tau, p_{jk}) = \begin{pmatrix}
-\frac{\partial^2 \ell(p_{jk}, \tau; t_j, t_k)}{\partial \tau_j \partial p_{jk}} \\
-\frac{\partial^2 \ell(p_{jk}, \tau; t_j, t_k)}{\partial \tau_k \partial p_{jk}}
\end{pmatrix}.
\]

\(h(p_{jk})\) is known as the score function associated with \(p_{jk}\), whereas \(h(p_{jk}, p_{jk}), H(\tau, \tau), H(\tau, p_{jk})\) are submatrices of the observed information matrix. The statistic in Equation 20 is asymptotically \(\chi^2\) distributed with one degree of freedom.

All necessary derivatives are given in the appendix. Observe that they are for the full log-likelihood function in Equation 9; that is, under the alternative bivariate lognormal model. However, when calculating Equation 20, the derivatives need to be evaluated not only at the MLEs of \(\tau_1\) and \(\tau_2\) in Equations 15 and 16 but also with \(p_{jk}\) set equal to its null value. For the \(H_0\) in Equation 18, this parameter thus vanishes from the derivatives, which simplify considerably. Besides, \(H(\tau, \tau)\) specializes to a diagonal matrix. As a result, the LM statistic in Equation 20 can easily be computed as follows:

\[
\text{LM}(p_{jk}) = \frac{\left(\sum_{i=1}^{n} \hat{\psi}_{ij} \hat{\psi}_{ik}\right)^2}{\sum_{i=1}^{n} \left(\hat{\psi}_{ij}^2 + \hat{\psi}_{ik}^2\right) - n - \left(\sum_{i=1}^{n} \hat{\omega}_i \hat{\psi}_{ij}\right)^2 + \sum_{i=1}^{n} \hat{\omega}_i^2},
\]

It is interesting to note the presence of the same key quantity in the numerators of \(\text{LM}(p_{jk})\) and the MLE of \(p_{jk}\) in Equation 17. As explained in Glas (1999), the
denominator of Equation 20 can be interpreted as the variance of the score function for the parameters under the null hypothesis adjusted for the estimation of the unknown parameters. It follows that $LM(\rho_{jk})$ can be interpreted as the ratio of the squared average cross-product of the residual logtimes and their adjusted variances, which explains its asymptotic behavior as a chi-square statistic.

4. Illustration With Empirical RTs

The effectiveness of the bivariate RT model in avoiding erroneous conclusions about collusion between test takers likely to arise when the inference is based on direct analyses of the RTs without any formal model is illustrated for a small sample of test takers from an administration of the Natural World Assessment (NAW-8) test in a study by Wise, Kong, and Pastor (2007). The NAW-8 is a test of quantitative and scientific reasoning proficiencies for college students. In an earlier study, we calibrated the same items under the lognormal RT model in Equation 4 and obtained excellent fit of the model to the items (Klein Entink et al., 2009).

In the Wise et al. (2007) administration of the NAW-8, the students served as participants in an experiment and had no stakes whatsoever in the test. Attempts to detect collusion between them would therefore be bound to fail. In fact, Wise et al. conducted their study to show the possibilities of using RTs for the detection of test takers with motivation problems. However, the goal of our current use of the data set is only to exemplify the proposed procedures for a few typical RT patterns. In particular, we wanted to show how focus on the estimation of correlation parameter ($\rho_{jk}$) in the model—or a statistical test of a hypothesis about it—can prevent from making the erroneous inferences suggested by a direct inspection of the RT patterns. Besides, the example is to illustrate the ease of calculating all relevant quantities in the procedures.

We analyzed the first 10 pairs of participants in the data set. The results for the three pairs of participants in Tables 1 and 2 were picked because they nicely illustrate the points we wanted to make with this example. The estimates of the time intensity and discrimination parameters of the items and the RTs of the participants in the example are shown in Table 1. The estimates of the item parameters were obtained from the full data set for 386 test takers using the identifiability restriction in Equation 5. Hence, the MLEs of the speed parameters of the six participants in the first row of Table 2, which were calculated directly from the logRTs and the $\alpha_i$ parameters of the items using Equations 15 and 16, center about zero. The next relevant quantities are the residual logRTs $\psi_{ij}$ in Equation 7, whose estimates follow directly once the MLEs of the speed parameters have been calculated. These estimates are also shown in Table 1. Remember that all model parameters are on the logarithmic scale, and that the same thus holds for the estimates of $\psi_{ij}$. 
A naive analysis of the agreement between the RTs of two test takers would focus on the Pearson correlation $r_{jk}$ between their RTs. For the three pairs of participants, these correlations in the second row of Table 2 seem to suggest substantial agreement between the behavior of the first two pairs and no agreement at all between the third pair. In addition, the lack of correlation for the third pair may seem to be consistent with their RT patterns in Table 1, which displays generally much greater RTs for Participant 5 (the difference in the size between the RTs of these two participants seems to be confirmed by the MLEs of their speed parameters: Participant 5 $[\hat{\tau}_5 = -0.53]$ worked considerably slower than Participant 6 $[\hat{\tau}_6 = 0.41]$).

However, the estimates of correlation parameter $\rho_{jk}$ in the bivariate RT model reveal an entirely different picture. The estimates for the first two pairs of participants point at negligible correlation: $\hat{\rho}_{12} = 0.02$ and $\hat{\rho}_{34} = -0.03$. The lack of correlation is confirmed by the values of the statistic for the test of $H_0$: $\rho_{jk} = 0$ in the last row of Table 2: $LM(\rho_{12}) = .02$ and $LM(\rho_{34}) = .04$ with $p_{12} = 90$ and $p_{34} = 83$. However, the estimate for the third pair of participants was negative: $\hat{\rho}_{56} = -0.36$. Its size was substantial enough to be surprising but too small...
to have much practical meaning: \( \text{LM}(\hat{r}_{56}) \) was equal to 1.71, with \( p_{56} = 0.19 \).

A check of the estimated residuals \( \hat{\psi}_{ij} \) for the third pair of participants in Table 1 reveals that the negative value of \( \hat{r}_{56} \) is almost entirely explained by extremely negative residuals for Participant 6 on only three of the items (items 3, 15, and 19).

### 5. Power of the Test

LM tests are optimal in several ways. For instance, they are known to be locally most powerful, consistent, and asymptotically equivalent to the better-known likelihood-ratio and Wald tests. The power of LM tests has been studied empirically for several applications in response modeling in the papers by Glas and his associates referred to earlier. To demonstrate the power of the current test, a simulation study was done for the items in the same NAW-8 test as in the previous section.

The power function of the test is the probability of rejecting the null hypothesis as a function of the true value of the parameters of interest, \( \rho_{jk} \). For a test of level \( \alpha \), the function is given by

\[
\Pr\{\text{LM}(\rho_{jk}) > \chi^2_{1-\alpha}(1) | \rho_{jk} \},
\]

where \( \chi^2_{1-\alpha}(1) \) is the \( (1 - \alpha) \)th quantile under a chi-square distribution with one degree of freedom. In this study, we used the conventional level of \( \alpha = 0.05 \).

Clearly, the power increases with the number of common items for the two test takers. We therefore simulated RTs under the bivariate lognormal model in Equation 8 for pairs of test takers on the first \( n = 20, 40, \) and \( 60 \) items of the NAW-8 and estimated the probability in Equation 26 as a function of \( \rho_{jk} = 0.00(.05).90 \). To check whether there would be any (unwanted) sensitivity of the test to the speed at which the test takers work, we simulated RTs at each of the true values of \( \rho_{jk} \) for each of the speed combinations...
For each combination of these parameter values, 1,000 replications were simulated. The resulting estimates of the power functions are given in Figure 1. As expected, each of the curves increased with the true value of $r_{jk}$. In addition, the curves for a longer test length dominated those for a shorter length. For $r_{jk} = 0$, the power of the test is always equal to its level—a feature that was reproduced by the results for each of the cases in Figure 1. In addition, across the four panels in Figure 1, the power functions appeared to be identical for all practical purposes, which indicates that the LM test was insensitive to the actual levels of speed of the test takers. Generally, the test appears to reach near perfect power at $r_{jk} = .4$ for test length $n = 60$, $r_{jk} = .6$ for $n = 40$, and $r_{jk} = .8$ for $n = 20$.

Just as for any other powerful statistical test, when using the proposed test in this article we should be aware of a possible embarrassment involved in having too much power. For instance, if for a test of $n = 40$ items the interest is not in detecting any correlations smaller than .6, it might be prudent to use the generalization of the test in the appendix, say, for $c = .4$. 

FIGURE 1. Estimated power functions of the Lagrange multiplier test of Equation 18 against Equation 19 for test lengths of $n = 20$ (dotted curve), 40 (dashed curve), and 60 (solid curve) and combinations of speed parameters $(\tau_j, \tau_k) = (0, 0), (-.1, .1), (-.2, .2)$ and $(-.3, .3)$. For each combination of these parameter values, 1,000 replications were simulated.
6. Concluding Remarks

The main point in this article is that RTs offer valuable information about the test takers’ behavior during a test, which can be profitably used to check data sets for suspicious agreement between the RT patterns of different test takers. However, to prevent capitalization on spurious relations, the RTs should be analyzed under a model for their joint distribution for the test takers.

The bivariate lognormal model proposed in this article captures the two main sources of spuriousness: the effects of the differences in time intensity between the items and the speed at which the test takers operate during the test. The first type of effect leads to correlation between the observed RTs that should not be confused with actual collusion between test takers. The first two pairs of participants in Table 2 illustrated this point. The second type leads to additional similarity between RTs when two test takers happen to work at approximately the same speed. For example, the agreement between the general level of the RTs between Participants 2, 3, and 5 is due to their generally working slowly on the test and does not point at any coordination between their behavior.

A third type of quantity that should always be checked, especially when the value of $\rho_{jk}$ need further explanation, are the estimated residuals $\hat{\psi}_{ij}$. The MLE of $\rho_{jk}$ appears to be just the observed correlation between these residuals; see Equation 17. We illustrated their use in our explanation of the negative value of $\rho_{jk}$ for the third pair of participants.

The examples also show how easy it is to conduct the proposed analysis. All relevant quantities have simple closed-form expressions that follow immediately from the item parameters and the RTs by the test takers. The only extra requirement is the calibration of the items under the RT model. However, this is also easily performed as part of the regular calibration of the test items (van der Linden, 2006).

Appendix

The necessary derivatives of the full log likelihood under the bivariate lognormal model in Equation 9 are given in the following two sections.

First-Order Derivatives

\[
\frac{\partial \ell(p_{jk}, \tau; t_j, t_k)}{\partial \tau_j} = \frac{1}{1 - \rho^2_{jk}} \sum_{i=1}^{n} \alpha_i (p_{jk} \psi_{ij} - \psi_{ij}),
\]

(A1)

\[
\frac{\partial \ell(p_{jk}, \tau; t_j, t_k)}{\partial \tau_k} = \frac{1}{1 - \rho^2_{jk}} \sum_{i=1}^{n} \alpha_i (p_{jk} \psi_{ij} - \psi_{ik}),
\]

(A2)

(continued)
Appendix (continued)

\[
\frac{\partial^2 \ell(\rho_{jk}, \tau; t_j, t_k)}{\partial \rho_{jk}^2} = \frac{n(1 + \rho_{jk}^2)}{(1 - \rho_{jk}^2)^2} \left\{ \sum_{i=1}^{n} \alpha_i^2 \right\} - \sum_{i=1}^{n} \frac{\psi_{ij}^2 - 2\rho_{jk} \psi_{ij} \psi_{ik} + \psi_{ik}^2}{(1 - \rho_{jk}^2)^2} + \frac{4\rho_{jk} \left( \psi_{ij}^2 - \rho_{jk}^2 \psi_{ij} \psi_{ik} + \rho_{jk} \psi_{ik}^2 - \psi_{ij} \psi_{ik} \right)}{(1 - \rho_{jk}^2)^3} \left\{ \sum_{i=1}^{n} \alpha_i^2 \right\} + \frac{2}{(1 - \rho_{jk}^2)^3} \left\{ \sum_{i=1}^{n} \alpha_i^2 \right\},
\]

\[
\frac{\partial^2 \ell(\rho_{jk}, \tau; t_j, t_k)}{\partial \tau_j \partial \tau_k} = \frac{\rho_{jk}}{(1 - \rho_{jk})} \sum_{i=1}^{n} \alpha_i^2,
\]

\[
\frac{\partial^2 \ell(\rho_{jk}, \tau; t_j, t_k)}{\partial \rho_{jk} \partial \tau_j} = \frac{(1 + \rho_{jk}^2) \sum_{i=1}^{n} \alpha_i \psi_{ij} - 2\rho_{jk} \sum_{i=1}^{n} \alpha_i \psi_{ij}}{(1 - \rho_{jk}^2)^2},
\]

\[
\frac{\partial^2 \ell(\rho_{jk}, \tau; t_j, t_k)}{\partial \rho_{jk} \partial \tau_k} = \frac{(1 + \rho_{jk}^2) \sum_{i=1}^{n} \alpha_i \psi_{ij} - 2\rho_{jk} \sum_{i=1}^{n} \alpha_i \psi_{ik}}{(1 - \rho_{jk}^2)^2}.
\]

Statistical Tests

The derivatives for the statistic for the test of \( H_0 : \rho_{jk} = 0 \) against \( H_1 : \rho_{jk} > 0 \) in Equations 20 through 24 follow on substitution of \( \rho_{jk} = 0 \) into Equations A1 through A9. This substitution simplifies the expressions considerably.

The LM statistic for the more general test of

\[ H_0 : \rho_{jk} = c \]
Appendix (continued)

against

\[ H_1 : p_{jk} > c. \]  \hspace{1cm} (A11) 

requires the use of the full derivatives for \( p_{jk} = c \). Both statistics have an asymptotic \( \chi^2 \) distribution with one degree of freedom.

References


**Author**

WIM J. VAN DER LINDEN is Chief Research Scientist at CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940; e-mail: wim_vanderlinden@ctb.com. His research interests include test theory, applied statistics, and research methods.

Manuscript received July 15, 2008
Revision received November 6, 2008
Accepted November 25, 2008