PSYCHOMETRIKA—VOL. 75, NO. 1, 120–139
MARCH 2010
DOI: 10.1007/S11336-009-9129-9

STATISTICAL TESTS OF CONDITIONAL INDEPENDENCE BETWEEN RESPONSES AND/OR RESPONSE TIMES ON TEST ITEMS

WIM J. VAN DER LINDEN

UNIVERSITY OF TWENTE AND CTB/MCGRAW-HILL

CEES A. W. GLAS

UNIVERSITY OF TWENTE

Three plausible assumptions of conditional independence in a hierarchical model for responses and response times on test items are identified. For each of the assumptions, a Lagrange multiplier test of the null hypothesis of conditional independence against a parametric alternative is derived. The tests have closed-form statistics that are easy to calculate from the standard estimates of the person parameters in the model. In addition, simple closed-form estimators of the parameters under the alternatives of conditional dependence are presented, which can be used to explore model modification. The tests were applied to a data set from a large-scale computerized exam and showed excellent power to detect even minor violations of conditional independence.

Key words: conditional independence, item-response theory (IRT), hierarchical modeling, Lagrange multiplier tests, lognormal model, response time.

1. Introduction

For a population of test takers, the responses to different test items typically correlate positively. Intuitively, this correlation may seem to make sense: a test taker who knows the answer to one item is likely to be more proficient than one who does not know it and should therefore have a higher probability of knowing the answer to the other item as well. However, as is well known, this argument confounds the correlation between the responses to test items with the impact of the proficiencies of the test takers. If we kept the proficiencies of the test takers constant—or in a more statistical language: condition on them—the argument would no longer hold, and the correlation between the responses would disappear. This is exactly what is postulated in the assumption of conditional (or "local") independence in item response theory (IRT) (e.g., Birnbaum, 1968; Lord, 1980).

Observe that this assumption of conditional independence supposes two different levels of randomness. At the level of a fixed test taker, the responses to test items are assumed to vary independently across replications. But at the level of a population of test takers, the proficiency also varies, and this variation creates the observed correlation between the lower-level responses. Examples of such hidden sources of covariation between observed variables can be found in almost any area of applied statistics. When one is identified, the correlation between the observed variables is usually referred to as "spurious correlation," and the change in the correlation upon conditioning on the covariate is known as Simpson's paradox.

This study received funding from the Law School Admissions Council (LSAC). The opinions and conclusions contained in this paper are those of the author and do not necessarily reflect the policy and position of LSAC.

Wim J. van der Linden is now at CTB/McGraw-Hill.

Requests for reprints should be sent to Wim J. van der Linden, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940, USA. E-mail: wim_vanderlinden@ctb.com

© 2009 The Psychometric Society. This article is published with open access at Springerlink.com

https://doi.org/10.1007/s11336-009-9129-9 Published online by Cambridge University Press

The same paradox may occur in response-time (RT) analyses. Instead of responses and proficiency, the two pertinent notions are now the RTs on the test items and the speed at which the test takers operate. The two are not the same; if two test takers respond to different test items and one produces a longer RT, it would be wrong to conclude that this person has worked at a slower speed. The solution to the item administered to this person may have required much more labor, and the person might actually have worked faster on it than the other person on the less-labor-intensive item.

It is helpful to notice an analogy between this notion of speed of cognitive labor and that of speed of motion in physics (van der Linden, 2009a). Without any reference to the distances traveled, it would be wrong to conclude that one car has moved faster than another only because it arrived earlier.

If test takers do differ in the speed at which they solve items, these differences serve as a potential source of covariation between the RTs on test items, and test takers with shorter RTs on one item can be expected to have shorter time on other items as well. Although we observe a positive correlation between pairs of items, the correlation will thus vanish as soon as we keep speed constant. Hence, it seems appropriate to assume conditional independence between RTs given speed as well.

An even more interesting relation exists between RTs and responses on a single item. Descriptive studies of RTs generally report negative correlations between them in the sense of longer RTs for incorrect than for correct solutions (e.g., Bergstrom, Gershon, & Lunz, 1994; Hornke, 2005, 2000; Swanson, Featherman, Case, Luecht, & Nungester, 1999; Swanson, Case, Ripkey, Clauser, & Holtman, 2001). This finding appeals to a general belief that test takers who do not know a solution, struggle longer and eventually are likely to settle on a wrong answer. These negative correlations are, however, entirely at odds with decades of experimental research on the speed-accuracy tradeoff in psychology, which have shown that, for a large variety of tasks, the likelihood of a correct solution tends to increase monotonically with time and thus that correctness and response time correlate positively (e.g., Luce, 1986, Section 6.5).

Usually, such conflicts between observed correlations point at hidden covariates as well. In the current case, the likely candidate is not proficiency or speed on their own but a higher-level relationship between them: If the two correlate positively in a population of test takers, the responses and RTs on a single item will tend to correlate negatively. A correct response is then likely to be the result of higher proficiency, and, as higher proficiency tends to be associated with higher speed, will be accompanied by a shorter RT.

On the other hand, a speed-accuracy tradeoff is a monotonically increasing relationship between speed and proficiency (i.e., a negative correlation) within a person. Therefore, experimental realization of the tradeoff will manifest itself as a positive correlation between time and correctness of the responses across different conditions of speed. But if speed and proficiency are held constant, these relationships between the two factors can no longer manifest themselves, and the responses and RTs should become independent. In sum, it also seems plausible to assume conditional independence between responses and RTs given proficiency and speed.

The preceding arguments were for a single test item (independence either between responses or RTs) or a pair of test items only (independencies between responses and RTs). The generalization to a full tests involves an additional assumption, namely that of constancy of speed and proficiency during the test. For instance, if constancy of speed was not guaranteed, any change of speed between two items would lead to a local violation of independence between the RTs on them. In fact, the speed-accuracy tradeoff suggests an accompanying change in proficiency and therefore a violation of the assumption of independence between the responses as well. However, as a test taker has better control of his/her speed than proficiency, we view constancy of speed as the more fundamental assumption of the two.

Of course, assumptions of constancy are idealizations; in real-world testing, speed and proficiency will always fluctuate somewhat, and violations of conditional independence should be

https://doi.org/10.1007/s11336-009-9129-9 Published online by Cambridge University Press

expected. These violations are no problem as long as they are minor and unsystematic. But larger, systematic violations may be indicative of more fundamental problems with the design of the test or the behavior of the test takers. Examples of possible design flaws are unrealistic time limits that force the test takers to speed up toward the end of test, fatigue because of an unduly large number of items in the test, and uncertainty about how to begin the test as the result of incomplete instructions.

In order to detect such flaws, we need statistical tools to check the validity of the conditional independence assumptions just identified. This paper offers two types of tools: formal statistical tests of these assumptions and estimates of relevant parameters under an alternative hypotheses of dependence that allow us to predict the impact of a modification of the assumption. The problem of statistical tests of the hypothesis of conditional independence between responses have received considerable earlier attention in the literature (e.g., Chen & Thissen 1997; Glas, 1999; Orlando & Thissen 2000; Yen, 1984), but the study of independence between RTs on different items and between responses and RTs on the same item is new.

In this research, we used the theory of Lagrange multiplier (LM) tests because it allowed us to follow an integrated approach to the three types of independence. Also, an advantage of the use of LM tests is the necessity to formulate specific parametric alternatives to the hypothesis of conditional independence. Equally important, LM statistics are generally easy to calculate; for the current set of hypotheses, they even turn out to be simple closed-form expressions based on standard estimates of the person parameters in the response or RT model. The proposed estimates of the alternative parameters, which allow us to explore the effects of modification of the independence assumptions, follow as a simple by-product of the LM statistics. For a general introduction to the LM test, which is equivalent to the Rao (1948) score test, see, for instance, Aitchison and Silvey (1958), Lehmann (1999, Section 7.7) or Silvey (1975, Section 7.4). LM tests have been used earlier to diagnose other violations of the fit of IRT models; for the current research, the results in Glas (1999), Glas and Dagohoy (2007), and Glas and Suárez Falcón (2003) are particularly important.

Obviously, the model we use as the null model to test the hypotheses of conditional independence has to be hierarchical with two different levels for the observations and dependencies between the parameters. Its first level consists of two distinct components, one for the distributions of the responses for a fixed person on a fixed item and the other for the distributions of the RTs. The second level has distinct components both for the joint distributions of the person and the item parameters in the first-level models. This type of hierarchical framework of modeling was proposed in van der Linden (2007). In the current research, we used the framework with precisely the component models suggested in this reference because these are well established and statistically fully tractable models. However, different specifications of the component models are certainly possible (e.g., more specific response models or models for polytomous items; exponential or Weibull models for the RTs instead of the lognormal model). For these alternative choices, the development of the statistical tests of conditional independence would have proceeded along the same lines.

2. Modeling Framework

On the first level, the Bernoulli distribution of response U_{ij} of a fixed person $j = 1, \ldots, N$ on a dichotomous item $i = 1, \ldots, n$ is assumed to be indexed by a probability of success that follows the well-known three-parameter logistic (3PL) model from IRT (Birnbaum, 1968; Lord, 1980). The probability is

$$P_{ij} = \Pr\{U_{ij} = 1; \theta_j; a_i, b_i, c_i\} \equiv c_i + (1 - c_i)\big\{1 + \exp\big[-a_i(\theta_j - b_i)\big]\big\}^{-1}, \tag{1}$$

https://doi.org/10.1007/s11336-009-9129-9 Published online by Cambridge University Press

where $\theta_j \in \mathbb{R}$ is the ability of test taker j, and $b_i \in \mathbb{R}$, $a_i \in \mathbb{R}^+$, and $c_i \in [0, 1]$ are the difficulty, discrimination, and guessing parameters for item i, respectively. Since θ_j is a person parameter that controls the probability of a correct response on the items, we will also refer to it as a parameter for the accuracy with which test taker j works on the items.

For the distribution of response time T_{ij} of test taker j on item i, we use a lognormal model,

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

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 intensity of item i, and $\alpha_i \in \mathbb{R}^+$ is its discrimination parameter. The sequel of this paper relies heavily on the fact that the model is equivalent to that of a normal distribution for the logarithm of the RT.

A key feature of the two models in (1)–(2) is their *separate* parameterization of the effects of the test taker and item on the response and RT distributions. In fact, except for the absence of a guessing parameters in the RT model, the two sets of parameters have analogous interpretations: θ_j and τ_j are parameters for the test taker's effect on the response and RT distributions, b_i and β_i represent the effects of the item on the locations of these distributions, and a_i and α_i control their variances. The next two components of the framework capitalize on these analogies.

The second-level population model describes the joint distribution of all first-level person parameters in the population of test takers as a bivariate normal distribution

$$(\theta_j, \tau_j) \sim \text{MVN}(\boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}) \tag{3}$$

with mean vector

$$\boldsymbol{\mu}_{\mathcal{P}} = (\mu_\theta, \mu_\tau) \tag{4}$$

and covariance matrix

$$\boldsymbol{\Sigma}_{\mathcal{P}} = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix}. \tag{5}$$

Likewise, for the distribution of the first-level item parameters in the domain of test items, we assume a multivariate normal distribution

$$(a_i, b_i, c_i, \alpha_i, \beta_i) \sim \text{MVN}(\boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}}) \tag{6}$$

with mean vector

$$\boldsymbol{\mu}_{\mathcal{I}} = (\mu_a, \mu_b, \mu_c, \mu_\alpha, \mu_\beta) \tag{7}$$

and covariance matrix

$$\boldsymbol{\Sigma}_{\mathcal{I}} = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{ac} & \sigma_{a\alpha} & \sigma_{a\beta} \\ \sigma_{ba} & \sigma_b^2 & \sigma_{bc} & \sigma_{b\alpha} & \sigma_{b\beta} \\ \sigma_{ca} & \sigma_{cb} & \sigma_c^2 & \sigma_{c\alpha} & \sigma_{c\beta} \\ \sigma_{\alpha a} & \sigma_{\alpha b} & \sigma_{\alpha c} & \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\beta a} & \sigma_{\beta b} & \sigma_{\beta c} & \sigma_{\beta\alpha} & \sigma_\beta^2 \end{pmatrix}. \tag{8}$$

In order to make the assumption of normality in (6) more plausible, some of the item parameters may have to be transformed first. In van der Linden (2007), it is suggested to use a log transformation for the discrimination parameters and a logit transformation for the guessing parameter for this purpose. The statistical tests of conditional independence presented below do not depend on the assumption of normality, though.

https://doi.org/10.1007/s11336-009-9129-9 Published online by Cambridge University Press

The modeling framework is not yet identifiable. A straightforward way of obtaining identifiability is to set

$$\mu_\theta = \mu_\tau = 0 \tag{9}$$

and

$$\sigma_\theta^2 = 1. \tag{10}$$

Bayesian methods of simultaneous estimation of all unknown parameters and model validation are given in Fox, Klein Entink, and van der Linden (2007), Klein Entink, Fox, and van der Linden (2009), and van der Linden (2007). The choice of lognormal models for RTs has a longer history. A lognormal model with an entirely different parameterization was presented in Thissen (1983). A restricted version of the model in (2) with $\alpha_i = \alpha$ was used to represent marginal RT distributions across test takers in an item bank by Schnipke and Scrams (1997). The current version of the model was proposed by van der Linden (2006); for statistical methods for estimating its parameters and evaluating its fit to empirical RTs, see this reference.

The lognormal model has been applied to detect differential speededness in multistage testing (van der Linden, Breithaupt, Chuah, & Zhang, 2007), to enhance test design (van der Linden, 2005, Section 9.5), to control speededness in adaptive testing (van der Linden, 2009b), and to detect aberrant test behavior (van der Linden & Guo 2008). The entire two-level framework can be exploited to increase IRT parameter estimation (van der Linden, Klein Entink, & Fox, 2008) as well as item selection in adaptive testing (van der Linden, 2008). Glas and van der Linden (2005) show that the combination of the two first-level models in (1) and (2) can be conceived of as an instance of a multidimensional IRT model for mixed response data.

3. Three Types of Conditional Independence

The first-level models have no common person or item parameters. This feature reflects the fact that the responses and RTs are assumed to be conditional independent. However, these models are linked through the covariance matrices $\boldsymbol{\Sigma}_{\mathcal{P}}$ and $\boldsymbol{\Sigma}_{\mathcal{I}}$ in (5) and (8). These second-level covariances allow for observed correlation between the responses and/or RTs of different test takers and/or items.

As already indicated, for the entire hierarchical framework, three different assumptions of conditional independence are entertained:

1. Independence between responses given θ. Formally, this assumption is defined as

$$f\left(u_{i_1}, \ldots, u_{i_G} \mid \theta\right) = \prod_{g=1}^{G} f\left(u_{i_g} \mid \theta\right) \tag{11}$$

   for $\theta \in \mathbb{R}$ and each possible subset of items of size $G \leq n$ with indices $(i_1, \ldots, i_G)$, where $f(u_{i_1}, \ldots, u_{i_G} \mid \theta)$ and $f(u_{i_g} \mid \theta)$ denote the probability functions of the responses on the subset of items and the individual items in it, respectively.
2. Independence between RTs given τ. This assumption is defined as

$$f\left(t_{i_1}, \ldots, t_{i_G} \mid \tau\right) = \prod_{g=1}^{G} f\left(t_{i_g} \mid \tau\right) \tag{12}$$

   for $\tau \in \mathbb{R}$ and each possible subset of items of size $G \leq n$ with indices $(i_1, \ldots, i_G)$, where $f(t_{i_1}, \ldots, t_{i_G} \mid \tau)$ and $f(t_{i_g} \mid \tau)$ are the densities of the RTs on the subset of items and the individual items in it, respectively.

https://doi.org/10.1007/s11336-009-9129-9 Published online by Cambridge University Press

3. Independence between responses and RTs given θ and τ. That is,

$$f(u_i, t_i \mid \theta, \tau) = f(u_i \mid \theta) f(t_i \mid \tau) \tag{13}$$

for $\theta, \tau \in \mathbb{R}$ and $i = 1, \ldots, N$.

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_N)$, $\mathbf{u}_j = (u_{1j}, \ldots, u_{nj})$, $\mathbf{t}_j = (t_{1j}, \ldots, t_{nj})$, $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_N)$, and $\mathbf{t} = (\mathbf{t}_1, \ldots, \mathbf{t}_N)$. Assuming that the items have been calibrated with sufficient precision to treat them as known, along with the standard assumptions of between-person independence of responses and RTs, the three independence assumptions imply the following product for the likelihood function of person parameters $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$:

$$f(\boldsymbol{\theta}, \boldsymbol{\tau} \mid \mathbf{u}, \mathbf{t}) = \prod_{j=1}^{N} \prod_{i=1}^{n} f(u_{ij} \mid \theta_j) f(t_{ij} \mid \tau_j). \tag{14}$$

Observe that for this full likelihood to factorize, the independence assumptions are required only for $G = n$. Nevertheless, checks on the conditional independence in real-world IRT applications typically are at the level of the adjacent pairs of items in the test. Although this choice is practically motivated, it does not seem to have any serious consequences. As indicated earlier, close relationships exist between violations of the assumptions of conditional independence and constancy of the speed and ability parameters during the test. If a test taker does not change speed between any pair of adjacent items in a test, and therefore the ability parameter remains constant as well, it seems safe to conclude to conditional independence at the level of all n items in the test.

4. Theory of LM Tests

The general approach we follow is to embed the models in (1) or (2) in a larger model with a plausible additional parameter that generates a violation of the independence assumption of concern. We then check the responses and RTs to see if the additional parameter can be assumed to vanish.

More generally, assume that the original model has parameters $\boldsymbol{\eta}_1$ and the alternative model additional parameters $\boldsymbol{\eta}_2$. The hypothesis we want to test is

$$\mathrm{H}_0 : \boldsymbol{\eta}_2 = \mathbf{0} \tag{15}$$

against

$$\mathrm{H}_1 : \boldsymbol{\eta}_2 \neq \mathbf{0}. \tag{16}$$

Let $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$. The statistic for an LM test of (15) against (16) is defined as

$$\mathrm{LM}(\boldsymbol{\eta}) = \mathbf{h}(\boldsymbol{\eta})' \mathbf{H}(\boldsymbol{\eta}, \boldsymbol{\eta})^{-1} \mathbf{h}(\boldsymbol{\eta}) \big|_{\boldsymbol{\eta}_1 = \widehat{\boldsymbol{\eta}}_1, \boldsymbol{\eta}_2 = \mathbf{0}}, \tag{17}$$

where $\mathbf{h}(\boldsymbol{\eta})$ is our generic notation for a score function, that is, the vector of first-order derivatives of the loglikelihood of the alternative model with respect to a vector of parameters $\boldsymbol{\eta}$,

$$\mathbf{h}(\boldsymbol{\eta}) = \frac{\partial}{\partial \boldsymbol{\eta}} \ln L(\boldsymbol{\eta}; \mathbf{x}). \tag{18}$$

https://doi.org/10.1007/s11336-009-9129-9 Published online by Cambridge University Press

Likewise, $\mathbf{H}(\boldsymbol{\eta}, \boldsymbol{\eta})$ is our notation for an observed information matrix with elements

$$h(\eta_p, \eta_q) = -\frac{\partial^2}{\partial \eta_p \partial \eta_q} \ln L(\boldsymbol{\eta}; \mathbf{x}). \tag{19}$$

Finally, $\widehat{\boldsymbol{\eta}}_1$ is the maximum-likelihood estimate (MLE) of $\boldsymbol{\eta}_1$, and $\mathbf{x}$ represents the data. The LM statistic is asymptotically χ^2 distributed with number of degrees of freedom equal to the dimension of $\boldsymbol{\eta}_2$ (e.g., Lehmann, 1999, Section 7.7; Silvey, 1975, Section 7.4).

Observe that the statistic in (17) is evaluated for the estimated model under the null hypothesis, $\mathrm{H}_0 : \boldsymbol{\eta}_2 = \mathbf{0}$. Hence, we only need to estimate $\boldsymbol{\eta}_1$. However, at the MLE of $\boldsymbol{\eta}_1$, $\mathbf{h}(\widehat{\boldsymbol{\eta}}_1) = \mathbf{0}$. These facts simplify the calculation of LM statistics enormously. The convenience of this feature will become clear in the applications of the test to the three different types of conditional independence below.

When the test is of a single parameter η_2 equal to zero, the LM statistic in (17) can be written as

$$\mathrm{LM}(\eta_2) = \left. \frac{h(\eta_2)^2}{h(\eta_2, \eta_2) - \mathbf{H}(\boldsymbol{\eta}_1, \eta_2)' \mathbf{H}(\boldsymbol{\eta}_1, \eta_1)^{-1} \mathbf{H}(\boldsymbol{\eta}_1, \eta_2)} \right|_{\boldsymbol{\eta}_1 = \widehat{\boldsymbol{\eta}}_1, \eta_2 = 0}. \tag{20}$$

To interpret (20), it is helpful to remember that, under the alternative model, the MLE of η_2 will also satisfy the likelihood equation, hence, $h(\widehat{\eta}_2) = 0$. Writing the numerator as $[h(\eta_2) - 0]^2|_{\eta_2 = 0}$, we are able to view this statistic as the standardized squared distance between the score function for the alternative parameter η_2 under the null model (i.e., at $\eta_2 = 0$) and at its estimated value under the alternative model (i.e., at $\eta_2 = \widehat{\eta}_2$). The standardizing factor can be shown to be equal to the asymptotic variance of $h_2(\eta_2)$ adjusted for the estimation of $\boldsymbol{\eta}_1$.

LM tests have nice statistical properties. For example, they are consistent, invariant under reparameterization, locally (i.e., for values of η_2 close to H_0) most powerful, and asymptotically equivalent to the likelihood-ratio (LR) and Wald tests (Lehmann, 1999, Sect. 7.7).

The tests presented in the next sections are for the case of items that have already been calibrated with enough precision to treat their item parameters (a_i, b_i, c_i) and (α_i, β_i) as known. This calibration is a prerequisite, for instance, for IRT-based test assembly and computerized adaptive testing. In a study of person fit under different polytomous IRT models by Glas and Dagohoy (2007), estimation of the item parameters had no noticeable impact neither on the type I error nor on the power of LM tests for different types of model violation. These results are assumed to hold generally for real-world applications of IRT models, which typically have many more responses per item than per test taker.

The only parameters with estimation error are thus the person parameters θ_j and τ_j. As a consequence, the expression for the LM statistic in (20) simplifies further because parameter vector $\boldsymbol{\eta}_1$ now also reduces to a scalar η_1.

4.1. Estimating the Alternative Parameters

Because LM tests are locally most powerful, with an increase in sample size, they will quickly tend to reject their null hypothesis for minor violations. This behavior should be welcomed. In our view, the primary purpose of the proposed tests is to identify which items can be treated as conforming to the model. Tests with higher power enable us to make such decisions with more reliance. On the other hand, rejection of one of the null hypotheses does not necessarily imply that the item is bad. The violation may just be minor but detectable because of the combination of larger sample sizes and the local power of the test close to the null value of the alternative parameter.

As a secondary step of analysis, for these flagged items, it is therefore helpful to have estimates of the values that their alternative parameters would take when they were left free. Such

https://doi.org/10.1007/s11336-009-9129-9 Published online by Cambridge University Press

estimates are common in other areas of applied statistics, e.g., linear covariance structure modeling (Sörbom, 1989), where they are used as indices to support the more explorative practice of model modification. But they can be used equally well for the models in this paper, as we will demonstrate in the empirical example below.

A standard approach is to use the result of one Newton–Raphson iteration for the alternative parameter η_2 from its value under H_0 as its estimator; that is, $\eta_2^{(0)} = 0$ minus the score function times the inverse of the Hessian both evaluated at this value. Or, more formally,

$$\widetilde{\eta}_2^{(1)} = \left\{ \eta_2^{(0)} + h(\eta_2^{(0)}) \frac{1}{h(\eta_2^{(0)}, \eta_2^{(0)}) - \mathbf{H}(\boldsymbol{\eta}_1, \eta_2^{(0)})' \mathbf{H}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_1)^{-1} \mathbf{H}(\boldsymbol{\eta}_1, \eta_2^{(0)})} \right\} \Bigg|_{\boldsymbol{\eta}_1 = \widehat{\boldsymbol{\eta}}_1, \eta_2^{(0)} = 0}. \quad (21)$$

The estimate has the same components as the LM statistic in (20) and is thus immediately available as a by-product of it.

Observe that (21) is not a maximum-likelihood estimate (MLE) of η_2. In particular, as the step is taken from the null value rather than an efficient initial estimate of η_2, it misses the (asymptotic) efficiency of the MLE. But in earlier empirical studies of IRT model fit (Glas, 1999), this one step-estimator yielded quite satisfactory approximations to the MLEs.

5. Test of Independence Between Responses

The alternative model is identical to that for a test of conditional independence for marginal maximum likelihood estimation of the item parameters in the 3PL model in Glas and Suárez Falcón (2003). Suppose that the test is for the pair of items (i, k). The alternative model is the conditional probability

$$\begin{aligned} P_{ikj} &= \Pr\{U_{ij} = 1; \theta_j, \delta_{ik} \mid U_{kj} = u_{kj}\} \\ &= c_i + (1 - c_i)\big[1 + \exp\big[-a_i(\theta_j - b_i - u_{kj}\delta_{ik})\big]\big]^{-1}. \end{aligned} \quad (22)$$

The model allows for different distributions of U_{ij} given $U_{kj} = 0$ and $U_{kj} = 1$, where the size of the differences between the two distributions depends on the value of additional parameter δ_{ik}. As the response probability in the regular model in (1) is equal to the expected value of the response by the test taker, new parameter δ_{ik} can be interpreted as a shift in the expected response on item i triggered by a correct response to item k by the same test taker.

Observe that the regular 3PL model follows for $\delta_{ik} = 0$. Therefore, for item pairs (i, k), we test the hypothesis

$$H_0 : \delta_{ik} = 0 \quad (23)$$

against

$$H_1 : \delta_{ik} \neq 0. \quad (24)$$

For test takers $j = 1, \ldots, N$ on items $i = 1, \ldots, n$ with response matrix $\mathbf{u} = (u_{ij})$, the log-likelihood of $(\boldsymbol{\theta}, \delta_{ik})$ is

$$\begin{aligned} \ell = \ell(\boldsymbol{\theta}, \delta_{ik}) &= \ln L(\boldsymbol{\theta}, \delta_{ik}; \mathbf{u}) \\ &= \sum_{j=1}^{N} \big[u_{ij} \ln P_{ikj} + (1 - u_{ij}) \ln(1 - P_{ikj})\big] \\ &\quad + \sum_{j=1}^{N} \sum_{\substack{l=1 \\ l \neq i}}^{n} \big[u_{lj} \ln P_{lj} + (1 - u_{lj}) \ln(1 - P_{lj})\big] \end{aligned} \quad (25)$$

https://doi.org/10.1007/s11336-009-9129-9 Published online by Cambridge University Press

with P_{lj} the regular response model defined in (1). For incomplete sampling designs, such as in computerized adaptive testing, not all test takers need to respond to the same items. If so, the sums are assumed to be taken only over the items actually taken.

By (20), the test statistic is

$$\text{LM}(\delta_{ik}) = \frac{h(\delta_{ik})^2}{h(\delta_{ik},\delta_{ik}) - \mathbf{H}(\boldsymbol{\theta},\delta_{ik})'\mathbf{H}(\boldsymbol{\theta},\boldsymbol{\theta})^{-1}\mathbf{H}(\boldsymbol{\theta},\delta_{ik})}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}},\delta_{ik}=0}, \tag{26}$$

where $\mathbf{H}(\boldsymbol{\theta},\boldsymbol{\theta})$ is an $N \times N$ diagonal matrix with elements $-\frac{\partial^2\ell}{\partial\boldsymbol{\theta}^2}$, and $\mathbf{H}(\boldsymbol{\theta},\delta_{ik})$ is a vector of length N with elements $-\frac{\partial^2\ell}{\partial\delta_{ik}\partial\theta_j}$. It follows that (26) can be rewritten as

$$\text{LM}(\delta_{ik}) = \frac{\left(\sum_{j=1}^N \frac{\partial\ell}{\partial\delta_{ik}}\right)^2}{\sum_{j=1}^N\left\{-\frac{\partial^2\ell}{\partial\delta_{ik}^2} + (\frac{\partial^2\ell}{\partial\delta_{ik}\partial\theta_j})^2(\frac{\partial^2\ell}{\partial\theta_j^2})^{-1}\right\}}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}},\delta_{ik}=0}. \tag{27}$$

Using the first- and second-order derivatives in Appendix A, the statistic simplifies to

$$\text{LM}(\delta_{ik}) = \frac{\left(\sum_{j=1}^N [u_{kj}\widehat{v}_{ij}(u_{ij} - \widehat{P}_{ij})]\right)^2}{\sum_{j=1}^N\left(-a_i u_{kj}\widehat{\zeta}_{ij} + \frac{(a_i u_{kj}\widehat{\zeta}_{ij})^2}{\sum_{l=1}^n a_i\widehat{\zeta}_{ij}}\right)}, \tag{28}$$

where $\widehat{P}_{ij}$ is the response probability in (1) evaluated at the MLE of θ_j, and $\widehat{v}_{ij}$ and $\widehat{\zeta}_{ij}$ are the expressions in (A.1) and (A.2) for $P_{ijk} = \widehat{P}_{ij}$. The statistic has an asymptotic χ^2 distribution with one degree of freedom.

Observe that the statistic is in closed form and that the estimates of θ required for it are part of the standard output of an IRT analysis.

6. Test of Independence Between RTs

As already noted, the RT model in (2) can be viewed as a normal density for $\ln T_{ij}$. This fact suggests using the following bivariate normal distribution of the logtimes on the pairs of items (i,k) as an alternative:

$$f\left(\ln t_{ij}, \ln t_{kj} \mid \tau_j, \rho_{ik}\right) = \frac{\alpha_i\alpha_k}{2\pi\sqrt{1-\rho_{ik}^2}}\exp\left\{\frac{-1}{2(1-\rho_{ik}^2)}\left(\psi_{ij}^2 - 2\rho_{ik}\psi_{ij}\psi_{kj} + \psi_{kj}^2\right)\right\}, \tag{29}$$

where

$$\psi_{ij} = \alpha_i\left[\ln t_{ij} - (\beta_i - \tau_j)\right]. \tag{30}$$

The alternative model has the extra parameter $|\rho_{ik}| \leq 1$, which is the correlation between the logtimes on items i and k by the same test taker. The same model has been proposed to use RTs for the detection collusion between pairs of test takers during a test (van der Linden, 2009c); for technical details, see this reference.

Obviously, under conditional independence, the correlation is equal to zero. Thus, for item pairs (i,k), the hypothesis to be tested is

$$\text{H}_0 : \rho_{ik} = 0 \tag{31}$$

https://doi.org/10.1007/s11336-009-9129-9 Published online by Cambridge University Press

against

$$\mathrm{H}_1 : \rho_{ik} \neq 0. \tag{32}$$

Under the null hypothesis, (29) factorizes into the product of the two densities for the RTs on item i and item k in (2).

For test takers $j = 1, \ldots, N$ on items $i = 1, \ldots, n$ with RT matrix $\mathbf{t} = (t_{ij})$, the loglikelihood of $(\boldsymbol{\tau}, \rho_{ik})$ can be written as

$$\begin{aligned}\ell(\boldsymbol{\tau}, \rho_{ik}) = \text{const} &- \frac{N}{2}\ln\left(1-\rho_{ik}^2\right) \\ &- \sum_{j=1}^{N} \frac{1}{2(1-\rho_{ik}^2)}\left(\psi_{ij}^2 - 2\rho_{ik}\psi_{ij}\psi_{kj} + \psi_{kj}^2\right) \\ &- \sum_{j=1}^{N}\sum_{\substack{l=1 \\ l\neq i,k}}^{n} \frac{1}{2}\psi_{lj}^2.\end{aligned} \tag{33}$$

In this application, the matrices $\mathbf{H}(\boldsymbol{\tau}, \boldsymbol{\tau})$ and $\mathbf{H}(\boldsymbol{\tau}, \rho_{ik})$ specialize in the same way as before. Therefore, test statistic LM (ρ_{ik}) is defined analogous to (27) with θ_j and δ_{ik} replaced by τ_j and ρ_{ik}, respectively. Using the derivatives in the Appendix A, the statistic can be written as

$$\mathrm{LM}(\rho_{ik}) = \frac{\left(\sum_{j=1}^{N} \widehat{\psi}_{ij}\widehat{\psi}_{kj}\right)^2}{\sum_{j=1}^{N}\left(\widehat{\psi}_{ij}^2 + \widehat{\psi}_{kj}^2 - 1 - \frac{(\alpha_k\widehat{\psi}ij + \alpha_i\widehat{\psi}kj)^2}{\sum_{i=1}^{n}\alpha_i^2}\right)}, \tag{34}$$

where

$$\widehat{\psi}_{ij} = \alpha_i\left[\ln t_{ij} - \left(\beta_i - \widehat{\tau}_j\right)\right]. \tag{35}$$

From (2) it is easy to verify that

$$\widehat{\tau}_j = \frac{\sum_{i=1}^{n}\alpha_i^2(\beta_i - \ln t_{ij})}{\sum_{i=1}^{n}\alpha_i^2} \tag{36}$$

is the MLE of τ_j. Substituting this estimate, (34) is simple to calculate.

7. Test of Independence Between Responses and RTs

We replace (13) by the equivalent independence assumption

$$f\left(t_{ij} \mid u_{ij}, \tau_j\right) = f\left(t_{ij} \mid \tau_j\right), \quad u_{ij} = 0, 1, \tag{37}$$

for all i and j. This form of the assumption is preferred over the alternate form $f(u_{ij} \mid t_{ij}, \theta_j) = f(u_{ij} \mid \theta_j)$, $t \in R$, for all i and j, for the following reasons: First, we only have to check the equality of the two conditional distributions of T_{ij} given $U_{ij} = 0$ and 1 instead of the equality of an entire family of distributions of U_{ij} given the continuous measure $T_{ij} = t_{ij}$. Second, for the same number of test takers, the estimation of location parameter τ_j in a normal distribution from continuous data is expected to be more accurate than the estimation of θ_j in an IRT model for binary data. Third, because the MLE of τ_j has the simple closed-form in (36), the statistic for the version of independence in (37) will be easier to calculate.

https://doi.org/10.1007/s11336-009-9129-9 Published online by Cambridge University Press

The alternative to the RT model in (2) is

$$f(t_{ij};\tau_j,\lambda_i)=\frac{\alpha_i}{t_{ij}\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\alpha_i\left(\ln t_{ij}-(\beta_i-\tau_j-u_{ij}\lambda_i)\right)\right]^2\right\}. \tag{38}$$

In this model, new parameter λ_i represents a shift in the location of the distribution of the RT on the item triggered by a correct response on it. As different RT distributions on an item given a correct and an incorrect response for the same test taker is the same thing as conditional dependence between responses and RTs, λ_i can be interpreted as a direct measure of it.

Thus, for item i, the hypothesis to be tested is

$$\mathrm{H}_0:\lambda_i=0 \tag{39}$$

against

$$\mathrm{H}_1:\lambda_i\neq 0. \tag{40}$$

The loglikelihood can be written as

$$\begin{aligned}\ell(\boldsymbol{\tau},\lambda_i)&=\ln\ell(\boldsymbol{\tau},\lambda_i;\mathbf{u}_i,\mathbf{t}_i)\\&=\text{const}-\frac{1}{2}\sum_{j=1}^{N}\xi_{ij}^2-\frac{1}{2}\sum_{j=1}^{N}\sum_{\substack{l=1\\l\neq i}}^{N}\xi_{lj}^2\end{aligned} \tag{41}$$

with

$$\xi_{ij}=\alpha_i\left[\ln t_{ij}-(\beta_i-\tau_j-u_{ij}\lambda_i)\right] \tag{42}$$

and

$$\xi_{lj}=\alpha_l\left[\ln t_{lj}-(\beta_l-\tau_j)\right]$$

for

$$l=1,\ldots,N\quad\text{and}\quad l\neq i.$$

Using the derivatives in Appendix A, from (27) the test statistic for the hypotheses in (39)–(40) follows as

$$\mathrm{LM}(\lambda_i)=\frac{\left(\sum_{j=1}^{N}u_{ij}\alpha_i\widehat{\xi}_{ij}\right)^2}{\sum_{j=1}^{N}\left(u_{ij}\alpha_i^2-\frac{(u_{ij}\alpha_i^2)^2}{\sum_i\alpha_i^2}\right)} \tag{43}$$

with $\widehat{\xi}_{ij}$ the estimate of ξ_{ij} evaluated at $\tau_j=\widehat{\tau}_j$ and $\lambda_i=0$, where $\widehat{\tau}_j$ is the MLE in (36).

8. Generalization to Higher-Order Independencies

As noted earlier, the tests of (23) and (31) against their alternatives are typically applied to check adjacent pairs of items, that is, with $k=i-1$. Although we expect any type of violation of conditional independence to manifest itself primarily at this level, the result of not having to reject any of these tests is necessary but not sufficient for the full assumptions in (11) and (12) to hold. This observation does not hold for the third type of conditional independence in (13), which has to be checked for the individual items only.

https://doi.org/10.1007/s11336-009-9129-9 Published online by Cambridge University Press

However, when violations of independencies between larger sets of consecutive items are to be checked, simultaneous versions of (26) and (34) are to be preferred over repeated applications of the tests for $k = i - 1$, $k = i - 2$, and so on. Such generalizations are now outlined.

For a triple of items (i, k, l), the alternative model in (22) can be generalized to

$$P_{iklj} = c_i + (1 - c_i)\{1 + \exp[-a_i(\theta_j - b_i - u_{kj}\delta_{ik} - u_{lj}\delta_{il})]\}^{-1}. \tag{44}$$

Let $\boldsymbol{\delta}_i = (\delta_{ik}, \delta_{il})$. In order to test the hypothesis $H_0 : \boldsymbol{\delta}_i = \mathbf{0}$, we need an LM statistic for parameter vector $\boldsymbol{\eta} = (\boldsymbol{\theta}, \boldsymbol{\delta}_i)$. Using the fact that we evaluate the statistic at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, (17) can now be written as

$$\mathrm{LM}(\boldsymbol{\delta}_i) = \mathbf{h}(\boldsymbol{\delta}_i)'\big[\mathbf{H}(\boldsymbol{\delta}_i, \boldsymbol{\delta}_i) - \mathbf{H}(\boldsymbol{\delta}_i, \boldsymbol{\theta})\mathbf{H}(\boldsymbol{\theta}, \boldsymbol{\theta})\mathbf{H}(\boldsymbol{\theta}, \boldsymbol{\delta}_i)\big]^{-1}\mathbf{h}(\boldsymbol{\delta}_i)\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}, \boldsymbol{\delta}_i=\mathbf{0}}, \tag{45}$$

where $\mathbf{H}(\boldsymbol{\theta}, \boldsymbol{\theta})$ is the same $N \times N$ diagonal matrix as above, but, analogous to (A.2), we now have ζ_{iklj} with P_{iklj} substituted for P_{ikj}. The matrices $\mathbf{H}(\boldsymbol{\delta}_i, \boldsymbol{\delta}_i)$ and $\mathbf{H}(\boldsymbol{\theta}, \boldsymbol{\delta}_i)$ have sizes 2×2 and $N \times 2$, respectively. The first-order and second-order derivatives of the loglikelihood with respect to δ_{il} and the mixed derivatives with respect to θ_j and δ_{il} in these matrices are entirely analogous to (A.4) and (A.7). The only new element is

$$\frac{\partial^2 \ell}{\partial \delta_{ik}\delta_{il}} = \frac{\partial^2 \ell(\theta_j, \delta_{ik}, \delta_{il}; u_{kj}, u_{lj})}{\partial \delta_{ik}\delta_{il}} = a_i u_{kj} u_{lj} \upsilon_{iklj} \zeta_{iklj} \tag{46}$$

(now also with υ_{iklj} instead of υ_{ikj}). Evaluation of (45) is therefore straightforward. The statistic has an asymptotic χ^2 distribution with two degrees of freedom.

The generalization of the test for the RTs is more involved. For a combination of G items, the alternative model in (29) becomes the multivariate normal

$$f(\ln \mathbf{t}_j; \tau_j, \boldsymbol{\Sigma}) = \frac{1}{2\pi |\boldsymbol{\Sigma}|^{-1/2}} \exp\left(-\frac{1}{2}\boldsymbol{\psi}'_j \boldsymbol{\Sigma}^{-1} \boldsymbol{\psi}_j\right) \tag{47}$$

with $\boldsymbol{\Sigma}$ a $G \times G$ covariance matrix with (known) diagonal elements α_i^{-2} and off-diagonal elements ρ_{ik}, and $\boldsymbol{\psi}_j$ a vector of size G with (30) as elements.

Let $\boldsymbol{\rho}$ be a vectorized form of the lower off-diagonal part of $\boldsymbol{\Sigma}$. To test the hypothesis H_0: $\boldsymbol{\rho} = \mathbf{0}$, we need the version of (45) for parameter vector $\boldsymbol{\eta} = (\boldsymbol{\tau}, \boldsymbol{\rho})$. However, for $G > 2$, it becomes more convenient to reparameterize (47) and work with the inverse of $\boldsymbol{\Sigma}$. (Observe that the null hypothesis implies a diagonal form of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$.) The elements of the matrices in (45) can now be derived from Lehmann (1999, Example 7.5.5). The statistic has an asymptotic χ^2 distribution with $G(G-1)/2$ degrees of freedom.

9. Empirical Example

An empirical study was conducted to see how well each of the tests behaved for a data set of $N = 1{,}104$ test takers on a large-scale computerized examination of 96 items. The examination had a multistage format with one routing test and second and third stages of two alternative 24-item subtests each. Prior to their earlier operational use, the items in the examination had been pretested and calibrated using the 3PL model in (1). In addition, in an earlier study, we used the RTs for the same sample of test takers on these items to calibrate them under the lognormal

https://doi.org/10.1007/s11336-009-9129-9 Published online by Cambridge University Press

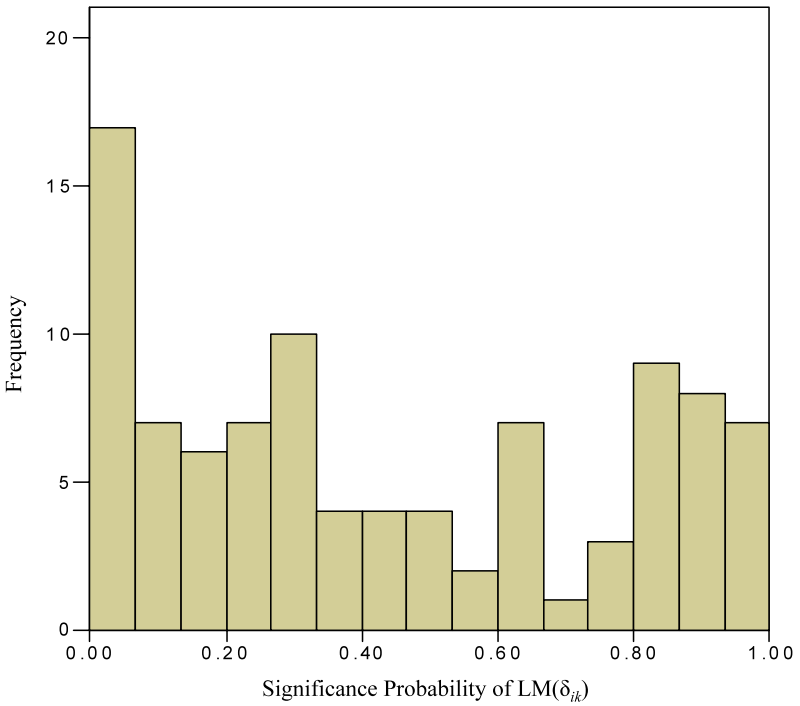


FIGURE 1.
Frequency distribution of significance probabilities of $\mathrm{LM}(\delta_{ik})$ $(n = 96)$.

model in (2). The model showed an excellent global fit to these RTs, except for a negligible tendency to shorter times in the lower tail of the distributions for some of the items (van der Linden et al., 2007, Figs. 2–4).

The test of the hypothesis of independence between responses and RTs in (39) was conducted for each of the 96 items involved in this study. Because the order of the items in the three subtests was randomized for each test taker, we conducted the other two tests only for the subsets of test takers who took a pair of items in the same order. More specifically, these tests were conducted for each of the 96 items in combination with the item that preceded it most frequently in the sample of test takers. The number of test takers for the pairs that were selected ranged from 24 to 79. This setup allowed us to use each individual item in this study. For each of these three cases, we also calculated the estimates of the alternative parameters in (21). As each of these statistics and estimates has a simple closed form with known quantities, they were easy to calculate.

Figures 1, 2, and 3 display the distributions of the significance probabilities of the $\mathrm{LM}(\delta_{ik})$, $\mathrm{LM}(\rho_{ik})$, and $\mathrm{LM}(\lambda_i)$ statistics in (28), (34), and (43) for the set of 96 items, respectively. The numbers of probabilities significant at the 5% level for the three tests were 17, 63, and 56, respectively. These results seem to suggest much larger numbers of violations of the assumption of conditional independence for the RTs as well as between the responses and the RTs relative to the assumption of independence between the responses only. Our next step should be to check these flagged items for the seriousness of their violations. When the violations are only minor, we know the hierarchical modeling framework with the current response and RT models still offers a useful explanation of the dependencies between these test items. Also, given the power of the test to detect these minor violations, it seems safe to assume that the quality of the remaining items in the test is entirely satisfactory.

https://doi.org/10.1007/s11336-009-9129-9 Published online by Cambridge University Press

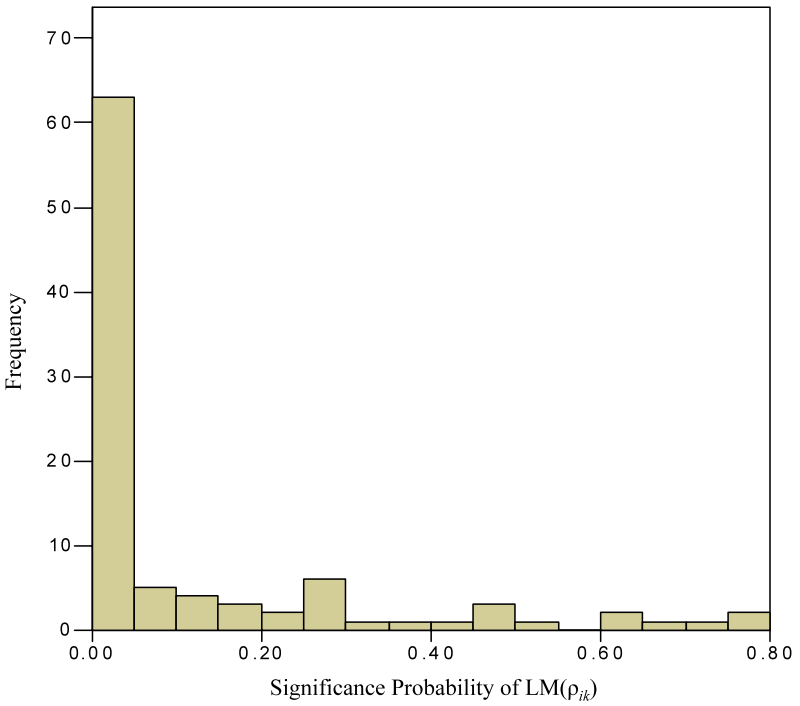


FIGURE 2.
Frequency distribution of significance probabilities of LM(ρ_{ik}) ($n = 96$).

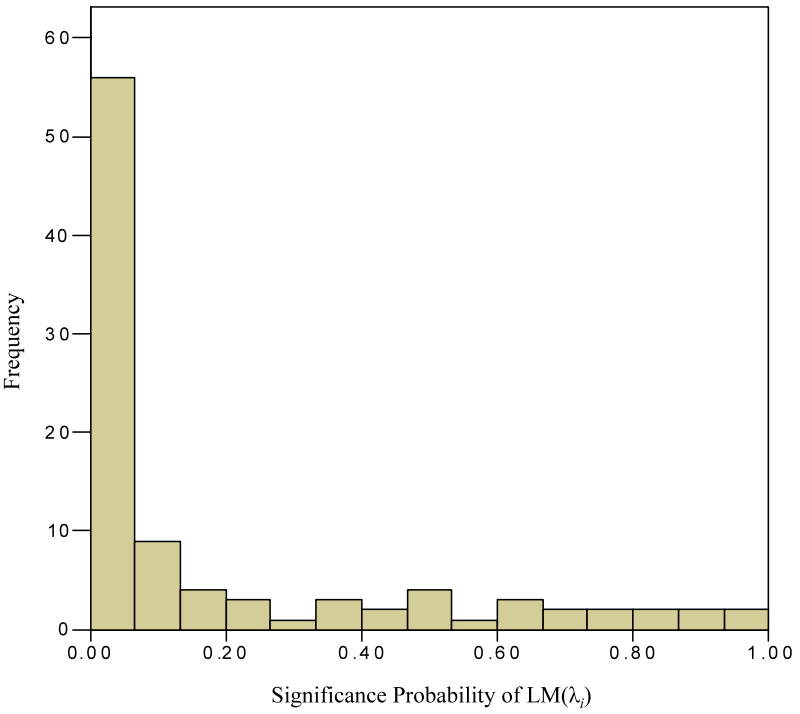


FIGURE 3.
Frequency distribution of significance probabilities of LM(λ_i) ($n = 96$).

https://doi.org/10.1007/s11336-009-9129-9 Published online by Cambridge University Press

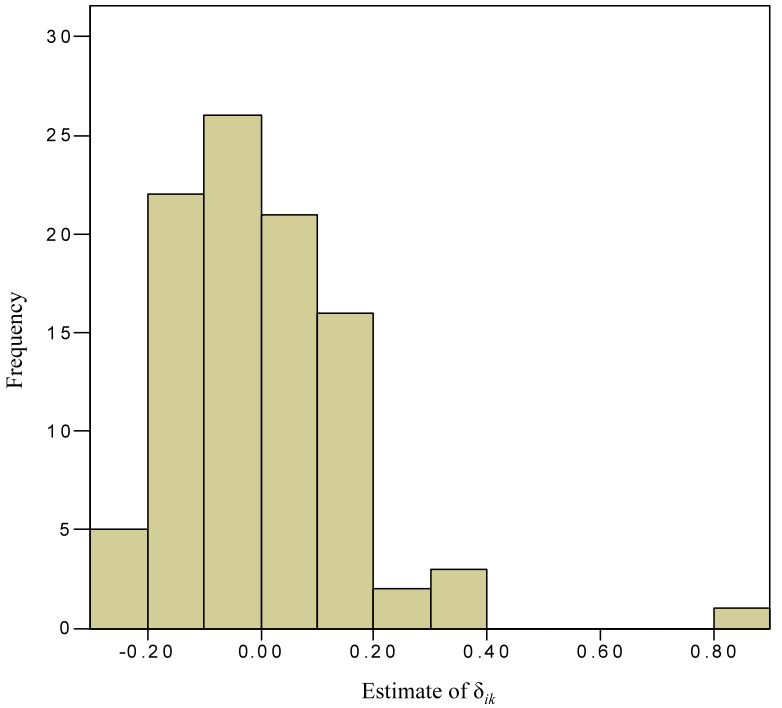


FIGURE 4.
Frequency distribution of estimates of δ_{ik} ($n = 96$).

Figures 4, 5, and 6 show the estimates of the alternative parameters in (21). The difference in sample size explains the large number of significant but negligible violations for LM(λ_i). This statistic was calculated across all 1,104 test takers in the data set, whereas the other two tests were only for the much smaller subsets of 24–79 test takers that responded to a common pair of items. As shown in Fig. 6, all estimates of λ_i center about zero with a standard deviation equal to 0.008. As λ_i is a parameter for the difference in the location of the logtime distributions between a correct and an incorrect response, the values for this parameter can be interpreted directly: Following (9), the average speed parameter for the test takers was set equal to zero. Besides, the average estimate of the time intensity parameter β_i of the items was 4.06. Hence, for an average test taker on an average item, a positive shift of a standard deviation of 0.008 in location would be equal to the difference between the values of 4.068 and 4.060 on the logarithmic scale, which is just 0.47 second on the regular time scale. (The average RT in the total data set was 75.39 seconds.)

Although the number of significant results for LM(ρ_{ik}) was also large, the average estimate of ρ_{ik} was only 0.06 with a standard deviation of 0.02 (see Fig. 5). These values do not seem to have any practical consequences either. The fact that the majority of the estimates was positive is consistent with a warm-up effect for the examination found in the earlier study of the same data set, in which a plot of the mean residual RTs against the administrative position of the items revealed that the test takers tended to operate slightly slower in the beginning and compensate later on in the examination (van der Linden et al., 2007, Fig. 7). Although systematic, the effect was quite minor, though: the difference between the mean residual RTs on the earlier and later items in the examination was approximately 1.7 seconds.

The distribution of the estimates of δ_{ik} centered about zero with a standard deviation equal to 0.16 (see Fig. 4), which suggests more serious violation of the assumption of independence

https://doi.org/10.1007/s11336-009-9129-9 Published online by Cambridge University Press

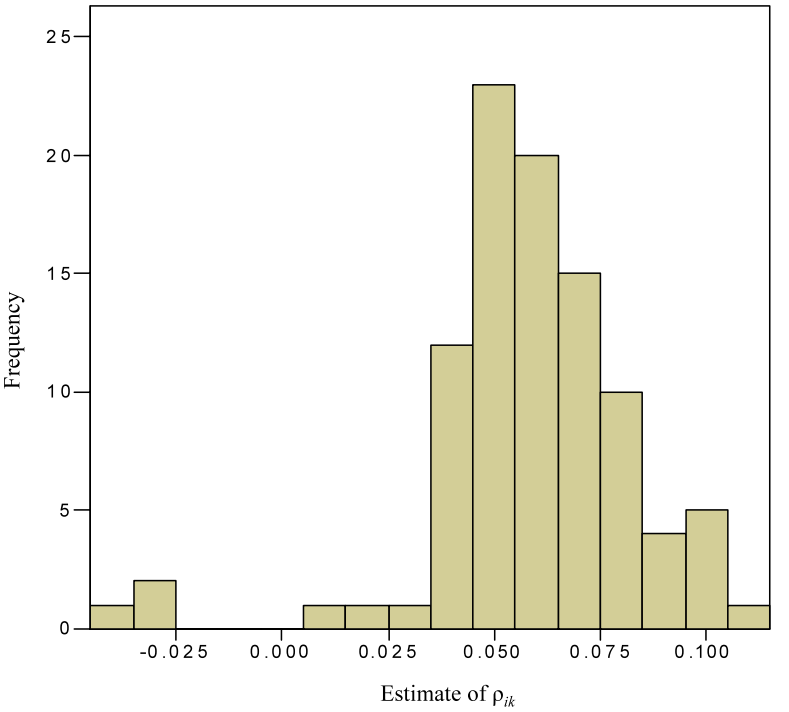


FIGURE 5.
Frequency distribution of estimates of ρ_{ik} ($n = 96$).

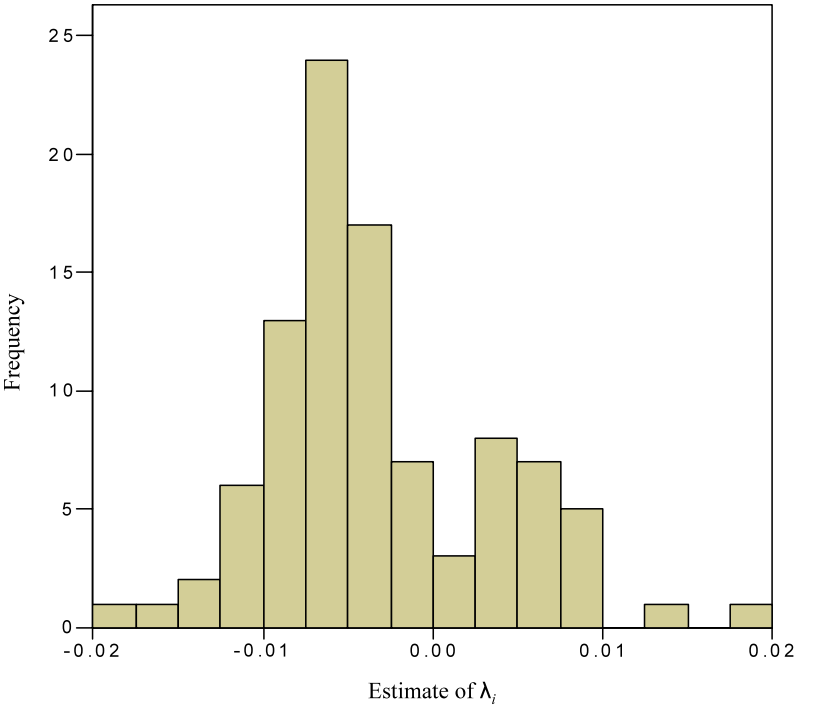


FIGURE 6.
Frequency distribution of estimates of λ_i ($n = 96$).

https://doi.org/10.1007/s11336-009-9129-9 Published online by Cambridge University Press

between the responses. Parameter δ_{ik} can be interpreted as a change in the probability of a correct response due to a previous correct response. Following (9), the average ability parameter for the test takers was also set equal to zero in this study. The average estimate of the item parameters for the model in (1) was: $\overline{a} = 0.597$, $\overline{b} = -0.297$, and $\overline{c} = 0.142$. Hence, for an average test taker on an item with these average parameter values, a value of δ_{ik} equal to one standard deviation (0.16) would mean a shift in the response probability from 0.533 to 0.567. This shift would still not be dramatic but should certainly raise more concerns than the shift for the RT distributions associated with the typical λ_i estimate above. In fact, for a few items, the estimates of δ_{ij} were much larger than 0.16. Figure 4 shows an estimate for one item that was even equal to 0.88. This estimate would certainly deserve closer inspection to find a reason for the violation of the independence assumption.

Because nearly all parameter estimates were generally small and irregular, we were unable to infer any systematic pattern between them. The correlations between the estimates were: $r_{\delta\rho} = -0.016$, $r_{\delta\lambda} = -0.023$, and $r_{\rho\lambda} = -0.095$.

10. Concluding

The goal of this research was to identify different plausible assumptions of conditional independence between responses and RT on test items. The assumptions are necessary for the hierarchical modeling framework in (1)–(8) to hold. Also, violations of these assumptions are indicative of potential design flaws in the test. The theory of Lagrange multiplier tests was used to derive formal statistical tests of the assumptions of conditional independence as well as easy-to-calculate estimates of the critical parameters under the alternative hypotheses of dependence that can be used to further diagnose the items. An empirical example showed how to use these statistical tools and suggested that, except for the violation of the assumption of conditional independent responses for an occasional item, all three assumptions were quite plausible.

The LM tests presented in this paper are not the only possibilities that may come to mind. For example, a standard test for the correlation in the bivariate normal distribution in (29) is Fisher's z test. In addition, a t test may seem attractive as a more conventional alternative for the shift in the normal distribution in (38). Finally, the statistical literature offers several tests of independence in a 2×2 table that may apply to the current case of conditional independence between dichotomous responses. But LM tests force us to be specific about the alternative hypothesis of dependence, have simple closed-form statistics, entail easy to calculate estimates of the alternative parameters, and have strong properties of optimality. Besides, it is attractive to be able to deal with all three hypothesis testing problems in the same statistical framework.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

Appendix A. First-Order and Second-Order Derivatives

For the first two sets of derivatives, the focus is on an item pair (i, k); for the third set, it is on a single item i.

https://doi.org/10.1007/s11336-009-9129-9 Published online by Cambridge University Press

A.1. Independence Between Responses

Let

$$\upsilon_{ikj} = \frac{a_i(P_{ikj} - c_i)}{(1 - c_i)P_{ikj}}, \tag{A.1}$$

$$\zeta_{ikj} = \upsilon_{ikj}\frac{1 - P_{ikj}}{(1 - c_i)}\left\{\frac{u_{ij}c_i}{P_{ikj}} - P_{ikj}\right\}. \tag{A.2}$$

The following derivatives of the loglikelihood in (25) are required (e.g., Lord, 1980):

$$\frac{\partial \ell}{\partial \theta_j} = \sum_{i=1}^{n} \upsilon_{ikj}(u_{ij} - P_{ikj}), \tag{A.3}$$

$$\frac{\partial \ell}{\partial \delta_{ik}} = -u_{kj}\upsilon_{ikj}(u_{ij} - P_{ikj}), \tag{A.4}$$

$$\frac{\partial^2 \ell}{\partial \theta_j^2} = \sum_{i=1}^{n} a_i \zeta_{ikj}, \tag{A.5}$$

$$\frac{\partial^2 \ell}{\partial \delta_{ik}^2} = a_i u_{kj}\zeta_{ikj}, \tag{A.6}$$

$$\frac{\partial \ell}{\partial \delta_{ik}\theta_j} = -a_i u_{kj}\zeta_{ikj}. \tag{A.7}$$

Since the statistic for the test of independence between a pair of responses will be evaluated at $\delta_{ik} = 0$, an attractive consequence is that its response probabilities do not depend on the response on item k but are just the probabilities in (1). Hence, in (28), P_{ij} can be substituted for P_{ikj}. Likewise, υ_{ij} and ζ_{ij} can be substituted for υ_{ikj} and ζ_{ikj}.

A.2. Independence Between RTs

Let

$$\psi_{ij} = \alpha_i\big[\ln t_{ij} - (\beta_i - \tau_j)\big]. \tag{A.8}$$

The following derivatives of the loglikelihood in (33) are required:

$$\frac{\partial \ell}{\partial \tau_j} = -\frac{1}{(1 - \rho_{ik}^2)}\big[\alpha_i\psi_{ij} - \rho_{ik}(\alpha_k\psi_{ij} + \alpha_i\psi_{kj}) + \alpha_k\psi_{kj}\big] - \sum_{\substack{l=1 \\ l\neq i,k}}^{n} \alpha_l\psi_{lj}, \tag{A.9}$$

$$\frac{\partial \ell}{\partial \rho_{ik}} = \frac{\rho_{ik} + \psi_{ij}\psi_{kj}}{1 - \rho_{ik}^2} - \frac{\rho_{ik}(\psi_{ij}^2 - 2\rho_{ik}\psi_{ij}\psi_{kj} + \psi_{kj}^2)}{(1 - \rho_{ik}^2)^2}, \tag{A.10}$$

$$\frac{\partial^2 \ell}{\partial \tau_j^2} = \frac{-\alpha_i^2 + 2\rho_{ik}\alpha_i\alpha_k - \alpha_k^2}{(1 - \rho_{ik}^2)} - \sum_{\substack{l=1 \\ l\neq i,k}}^{n} \alpha_l^2, \tag{A.11}$$

$$\frac{\partial^2 \ell}{\partial \rho_{ik}^2} = \frac{1 + \rho_{ik}^2 - \psi_{ij}^2 + 6\rho_{ik}\psi_{ij}\psi_{kj} - \psi_{kj}^2}{(1 - \rho_{ik}^2)^2} - \frac{4\rho_{ik}^2(\psi_{ij}^2 - 2\rho_{ik}\psi_{ij}\psi_{kj} + \psi_{kj}^2)}{(1 - \rho_{ik}^2)^3}, \tag{A.12}$$

$$\frac{\partial^2 \ell}{\partial \tau_j \partial \rho_{ik}} = \frac{\alpha_k\psi_{ij} + \alpha_i\psi_{kj}}{(1 - \rho_{ik}^2)} - \frac{2\rho_{ik}[\alpha_i\psi_{ij} - \rho_{ik}(\alpha_k\psi_{ij} + \alpha_i\psi_{kj}) + \alpha_k\psi_{kj}]}{(1 - \rho_{ik}^2)^2}. \tag{A.13}$$

https://doi.org/10.1007/s11336-009-9129-9 Published online by Cambridge University Press

The statistic for the test of independence between pairs of RTs will be evaluated at $\rho_{ik} = 0$. As a result, the derivatives simplify considerably.

A.3. Independence Between Responses and RTs

Let

$$\xi_{ij} = \alpha_i\big[\ln t_{ij} - (\beta_i - \tau_j - u_{ij}\lambda_i)\big]. \tag{A.14}$$

The following derivatives of the loglikelihood in (41) are required:

$$\frac{\partial \ell}{\partial \tau_j} = -\alpha_i \xi_{ij}, \tag{A.15}$$

$$\frac{\partial \ell}{\partial \lambda_i} = -\alpha_i u_{ij} \xi_{ij}, \tag{A.16}$$

$$\frac{\partial^2 \ell}{\partial \tau_j^2} = -\alpha_i^2, \tag{A.17}$$

$$\frac{\partial^2 \ell}{\partial \lambda_i^2} = -\alpha_i^2 u_{ij}, \tag{A.18}$$

$$\frac{\partial^2 \ell}{\partial \tau_j \partial \lambda_i} = -\alpha_i^2 u_{ij}. \tag{A.19}$$

The statistic for the test of independence between a responses and an RT will be evaluated at $\lambda_i = 0$. Hence, ξ_{ij} reduces to the argument of the regular lognormal model, and the LM statistic simplifies considerably.

References

Aithchison, J., & Silvey, D.C. (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics*, *29*, 813–828.

Bergstrom, B., Gershon, R., & Lunz, M.E. (1994). *Computer-adaptive testing: exploring examinee response time using hierarchical linear modeling*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading: Addison-Wesley.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289.

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

Glas, C.A.W. (1999). Modification indices for the 2PL and the nominal response model. *Psychometrika*, *64*, 273–294.

Glas, C.A.W., & Dagohoy, A.V.T. (2007). Person fit tests for IRT models for polytomous items with estimated person and item parameters. *Psychometrika*, *72*, 159–180.

Glas, C.A.W., & Suárez Falcón, J.C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*, 87–106.

Glas, C.A.W., & van der Linden, W.J.. (2005). *Likelihood-based estimation methods for models for concurrent continuous and discrete responses* (LSAC Report). Enschede, The Netherlands: University of Twente, Department of Research Methodology, Measurement, and Data Analysis.

Hornke, L.F. (2000). Item response times in computerized adaptive testing. *Psicológica*, *21*, 175–189.

Hornke, L.F. (2005). Response time in computer-aided testing: a "Verbal Memory" test for routes and maps. *Psychological Science*, *2*, 280–293.

Klein Entink, R.H., Fox, J.-P., & van der Linden, W.J. (2009). A multivariate multilevel approach to simultaneous modeling of accuracy and speed on test items. *Psychometrika*, *74*, 21–48.

Lehmann, E.L. (1999). *Elements of large-sample theory*. New York: Springer.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Erlbaum.

Luce, R.D. (1986). *Response times: their roles in inferring elementary mental organization*. Oxford: Oxford University Press.

https://doi.org/10.1007/s11336-009-9129-9 Published online by Cambridge University Press

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50–64.
Rao, C.R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society*, *44*, 50–57.
Schnipke, D.L., & Scrams, D.J. (1997). *Representing response time information in item banks* (LSAC Computerized Testing Report No. 97-09). Newtown, PA: Law School Admission Council.
Silvey, S.D. (1975). *Statistical inference*. London: Chapman & Hall.
Sörbom, D. (1989). Model modification. *Psychometrika*, *54*, 371–384.
Swanson, D.B., Featherman, C.M., Case, S.M., Luecht, R.M., & Nungester, R. (1999). *Relationship of response latency to test design, examinee proficiency and item difficulty in computer-based test administration*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.
Swanson, D.B., Case, S.E., Ripkey, D.R., Clauser, B.E., & Holtman, M.C. (2001). Relationships among item characteristics, examinee characteristics, and response times on USMLE Step 1. *Academic Medicine*, *76*, 114–116.
Thissen, D. (1983). Timed testing: an approach using item response theory. In D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.
van der Linden, W.J. (2005). *Linear models for optimal test design*. New York: Springer.
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. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *32*, 5–20.
van der Linden, W.J. (2009a). Conceptual issues in response-time modeling. *Journal of Educational Measurement, 46*. In press.
van der Linden, W.J. (2009b). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement*, *33*, 25–41.
van der Linden, W.J. (2009c). A bivariate lognormal response-time model for the detection of collusion between test takers. *Journal of Educational and Behavioral Statistics, 34*. In press.
van der Linden, W.J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*, 365–384.
van der Linden, W.J., Breithaupt, K., Chuah, D., & Zhang, O. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, *44*, 117–130.
van der Linden, W.J., Klein Entink, R.H., & Fox, J.-P. (2008). IRT parameter estimation with response times as collateral information. Manuscript submitted for publication.
Yen, W.M. (1984). Effects of local independence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125–145.

Manuscript Received: 23 JUL 2008
Final Version Received: 23 APR 2009
Published Online Date: 29 MAY 2009

https://doi.org/10.1007/s11336-009-9129-9 Published online by Cambridge University Press