PSYCHOMETRIKA—VOL. 68, NO. 2, 251–265
JUNE 2003

USING RESPONSE TIMES TO DETECT ABERRANT RESPONSES IN COMPUTERIZED ADAPTIVE TESTING

WIM J. VAN DER LINDEN AND EDITH M.L.A. VAN KRIMPEN-STOOP

UNIVERSITY OF TWENTE

A lognormal model for response times is used to check response times for aberrances in examinee behavior on computerized adaptive tests. Both classical procedures and Bayesian posterior predictive checks are presented. For a fixed examinee, responses and response times are independent; checks based on response times offer thus information independent of the results of checks on response patterns. Empirical examples of the use of classical and Bayesian checks for detecting two different types of aberrances in response times are presented. The detection rates for the Bayesian checks outperformed those for the classical checks, but at the cost of higher false-alarm rates. A guideline for the choice between the two types of checks is offered.

Key words: aberrant response patterns, computerized adaptive testing, posterior predictive checks, person misfit, residual analysis, response times.

Though the primary use of response vectors in testing is to construct accurate ability estimates, they also contain useful information to detect possible aberrances in examinee behavior. Most statistical analyses to identify such behavior belong to the class of statistical procedures known as residual analysis. That is, they are based on the residuals of an examinee's response vector left after a model known to explain the responses of a population of regular examinees has been fitted. The first step in this residual analysis is to check for examinees with unexpected behavior. A more challenging second step is to diagnose their response vector for specific types of aberrances. Ideally, the analysis would support the hypothesis of one type of aberrance and exclude competing hypotheses. Papers with seminal techniques for this type of analysis are Bradlow, Weiss, and Cho (1998), Drasgow, Levine and Williams (1985), Levine and Rubin (1979), Molenaar and Hoijtink (1990), and Trabin and Weiss (1983). A review of the literature is given in Meijer and Sijtsma (1995).

The introduction of computerized adaptive testing (CAT) has increased the necessity of checks on examinee behavior. For example, a new type of aberrance is response behavior due to preknowledge of some of the items in the pool. To return investments, item pools in CAT have to remain operational for some time. Particularly in high-stakes testing programs, examinees may try to use this time to memorize and share items in the pool. Another potential new source of aberrant behavior is due to differential speededness of the test. Adaptive tests are selected to have optimal information at the ability level of the examinee. However, items differ not only in their information but also in the amount of time they require. As a consequence, some CAT examinees operate under higher time pressure than others. A recent study revealed that high-ability examinees may suffer especially from this type of speededness. For those examinees, item selection CAT results in more difficult items, and more difficult items generally require more time (van der Linden, Scrams & Schnipke, 1999).

Also, residual analysis of response vectors has difficulty maintaining its power when applied to adaptive tests. One reason is that adaptive tests are typically much shorter than paper-and-

This study received funding from the Law School Admission Council (LSAC). The opinions and conclusions contained in this paper are those of the authors and do not necessarily reflect the policy and position of LSAC. The authors are most indebted to Wim M. M. Tielen for his computational assistance and to the US Defense Manpower Data Center for the permission to use the ASVAB data set in the empirical examples.

Requests for reprints should be sent to W.J. van der Linden, Department of Research Methodology, Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, THE NETHERLANDS. E-Mail: w.j.vanderlinden@utwente.nl

0033-3123/2003-2/2001-0874-A $00.75/0

pencil tests. More importantly, in an adaptive test the difficulty of the items converge to the examinee's ability level. Hence, items toward the end of the test tend to have a probability of success close to .50. However, residual analysis typically has its maximum power for tests with probabilities of correct responses close to one or zero. For more details on these issues, see van Krimpen-Stoop and Meijer (1999, 2000).

One possible way to counter these problems is to complement analysis of response data with analyses of other types of data. An obvious additional source of data are the examinee's response times. In computerized testing the time between keystrokes can easily be stored. Particularly in high-stake testing, the assumption that these records reflect the time the examinee actually needed to process the items and produce a response seems realistic. More importantly, unexpected response times can be indicative of specific types of aberrant behavior. For example, examinees who know some of the items prior to the test may answer them more quickly than generally required, while examinees running out of time will tend to show a series of unexpected short response times toward the end of the test.

There are also two more technical advantages involved in the use response times to check examinee behavior for aberrances. First, a major obstacle in residual analysis of item responses is their discrete nature. Especially for statistics that transform residuals to an asymptotically normally distributed variable, application to single dichotomous responses can hardly be recommended. Response times are continuous and do not possess this disadvantage. Second, as already noted, response patterns in CAT are the result of a built-in tendency to probabilities of success close to .50 toward the end of the test. This tendency implies residuals with values $-.50$ or .50, which are equally likely. Tendencies to such constrained distributions of residuals do not exist for response times.

To detect aberrancies in response times, a model with separate parameters for the items and the examinee is needed. In this paper, response times are modeled as a variable with a lognormal distribution. At the same time, the responses to the items are assumed to follow a 3-parameter logistic (3PL) response model. If both models fit, responses and response times for a given examinee are independent. Test statistics defined on these variables are thus also independent.

Both classical and Bayesian (posterior predictive) checks are presented. As usual, the classical statistics involve large-sample approximations that sometimes are not realistic. The Bayesian statistics involve the (mild) assumption of a prior distribution but produce exact distributions.

IRT and Response-Time Model

The item pool is indexed by $i = 1, \ldots, I$, whereas the items in the adaptive test are indexed by $k = 1, \ldots, n$. Thus, i_k is the i-th item in the pool administered as the k-th item in the test. The double level of subscripting is introduced for the analyses below where we focus on one item in the test and use parameter estimates based on the responses and response times on the other items. Examinee j's responses and response times on the items in the test are denoted as random vectors $\mathbf{U}_j = (U_{i_1 j}, \ldots, U_{i_n j})$ and $\mathbf{T}_j = (T_{i_1 j}, \ldots, T_{i_n j})$, with realizations $\mathbf{u}_j = (u_{i_1 j}, \ldots, u_{i_n j})$ and $\mathbf{t}_j = (t_{i_1 j}, \ldots, t_{i_n j})$, respectively.

In the empirical examples below, a 3PL model was fitted to the items $i = 1, \ldots, I$:

$$p_i(\theta) = \Pr\{U_i = 1\} \equiv c_i + (1 - c_i)\frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}, \tag{1}$$

where U_i is the response variable for item i, with $U_i = 1$ for a correct and $U_i = 0$ for an incorrect response, $\theta \in R$ is the ability of the examinee, and $a_i \in (0, \infty)$, $b_i \in R$, and $c_i \in [0, 1)$ are the discrimination, difficulty, and guessing parameter for item i, respectively.

In addition, the following loglinear model was fitted to the response times:

$$\ln T_{ij} = \mu + \delta_i + \tau_j + \varepsilon_{ij} \tag{2}$$

with

$$\varepsilon_{ij} \sim N(o, \sigma^2), \tag{3}$$

where δ_i is a parameter for the response time required by item i, τ_j is a parameter for the slowness of examinee j, μ is a parameter indicating the general response time level for the population of examinees and pool of items, and ε_{ij} a normally distributed residual or interaction term for item i and examinee j with mean 0 and variance σ^2. It follows that

$$\ln T_{ij} \sim N(\mu + \delta_i + \tau_j, \sigma^2). \tag{4}$$

A loglinear model for response times has been used earlier by Thissen (1983), whereas the parameterization in (4) is discussed Schnipke and Scrams (1997).

Observe that the model in (4) is formulated at the level of an individual examinee and item with ε_{ij} as the only random variable. Hence, $\ln T_{ij}$ and ε_{ij} have the same variance. The variance is modeled to be a common quantity across items and examinees, that is, $\sigma_{ij}^2 = \sigma^2$. A more general model is possible relaxing this assumption, but the current version is more parsimonious and showed excellent fit to the data set used in the empirical example below. Also, observe that, through the presence of the μ parameter in the model, all other parameters are scaled to have expectations zero. In particular, it holds that

$$\mathrm{E}(\tau) = 0. \tag{5}$$

Alternative Models and Approaches

The models in (1) and (4) are formulated for marginal distributions of responses and response times. The question can be raised if a model for a joint distribution would not be more realistic, because ability and slowness could very well be dependent on each other. However, it should be noted that, entirely in the tradition of test theory, either of the two models are at the level of a fixed examinee and a fixed item. If they fit, the distributions of the responses and response times are independent, and it is not necessary to model the joint distribution. The 3PL model has a proven history of satisfactory fit in testing. The only thing the response-time model implies is that if we know the amount of time the item requires and the slowness of this examinee, we are able to predict the response time for the item and examinee up to a random error. If an interaction term were added to (4), the model would become fully saturated, and a check on the goodness of fit of the model would amount only to a check on the assumption of the lognormal error distribution in (3). For the data set in the application below, the lognormal model was tested against models based the normal, Gamma, and Weibull distributions and showed excellent fit.

It is also possible to model the slowness parameter as a function of the ability parameter. For a correctly specified function, we would have fewer parameters to estimate during the adaptive test and can therefore expect increased efficiency of estimation. However, it is easy to misspecify such a function, and then its presence would bias the estimates of these parameters. For example, we may expect a monotonically decreasing relation between slowness and ability ("more able students work faster"). This assumption was introduced in a response time model by Thissen (1983). On the other hand, the existence of the speed-accuracy paradox in the literature on response latencies seems to imply a monotonically increasing relation: When tested, examinees typically have the option to choose between speed and accuracy. If they choose for accuracy, both the values for the ability parameter in (1) and the slowness parameter in (2) go up. Finally, there is also empirical information that the relation between ability and slowness depends on the level of speededness of the test (Swanson, in van der Linden, Scram & Schnipke, 1999). Given the fact that our insight in the mechanisms underlying response latencies is still incomplete, we have refrained from the option of modeling one examinee parameter as a function of the other.

Another possibility is a multilevel approach for a *population* of examinees, with the models in (1) and (4) as first-level models for the examinees and a second level model for the (joint)

distribution of their parameters. This approach would allow us to capture the correlation between examinee parameters θ and τ in the population (van der Linden, 2002). If an informative estimate of this correlation is available, estimation of θ and τ during the adaptive test can be made more efficient by using an empirical Bayes approach with a joint prior that reflects the correlation in the population. However, in our application, we found negligible empirical correlation between θ and τ(.035). We therefore used a Bayesian approach with marginal priors for these parameters. (We did find a correlation of .65 between item difficulty and response time parameters b_i and δ_i, though. But this information was useless because, as is usual in real-life adaptive testing, large-sample estimates of these parameters were treated as their true values.)

Parameter Estimation

The values of the parameters in (1) for the items in the pool are assumed to be estimated accurately enough to consider them as known during operational testing. The value of examinee j for θ is estimated using the expected a posterior (EAP) estimator. Let $f(\theta)$ be the prior density function of θ. For this prior the posterior distribution of θ has density function

$$f(\theta \mid u_{i_1 j}, \ldots, u_{i_n j}) \propto f(\theta) \prod_{k=1}^{n} p_{i_k}(\theta)^{u_{i_k j}} [1 - p_{i_k}(\theta)]^{1-u_{i_k j}}. \tag{6}$$

The EAP estimator of θ is defined as

$$\widetilde{\theta}_j \equiv \int \theta f(\theta \mid u_{i_1 j}, \ldots, u_{i_n j})\, d\theta. \tag{7}$$

The likelihood in (6) is not the full likelihood because it does not model the item-selection mechanism in the adaptive tests that selects the next item as a function of the observed responses on the previous items. However, ignoring the item-selection mechanism is correct if the interest is in Bayesian or direct likelihood inference about θ (Mislevy & Chang, 2000; see also Mislevy & Wu, 1966).

The response-time model in (4) has item parameters, $\delta_i, i = 1, \ldots, I$, and structural parameters, μ and σ^2. The values of these parameters can be estimated along with those for the item parameters in the response model during item calibration. As the model involves a linear decomposition for the location of a normally distributed variable, these values are estimated easily following an analysis-of-variance approach. For more details on the estimation of the item and structural parameters, see van der Linden, Scrams, and Schnipke (1999). Again, it is assumed that the values of these parameters have been estimated with enough accuracy to consider them as known.

In addition, the model has examinee parameters, τ_j, with values that can be estimated from the actual response times by the examinees during the test. For a vector of response times $(T_{i_1 j}, \ldots, T_{i_n j})$, the maximum likelihood estimator (MLE) of τ_j is:

$$\widehat{\tau}_j = \frac{\sum_{p=1}^{n} (\ln T_{i_p j} - \delta_{i_p})}{n} - \mu. \tag{8}$$

Observe that the terms in (8) are independent and normally distributed with common variance σ^2/n^2 and expectation τ_j/n. Therefore, this estimator has expectation

$$\mathrm{E}(\widehat{\tau}_j) = \tau_j \tag{9}$$

and variance

$$\mathrm{Var}(\widehat{\tau}_j) = \frac{\sigma^2}{n}. \tag{10}$$

A Bayesian approach to parameter estimation uses the posterior distribution of τ_j given the actual response times $(t_{i_1 j}, \ldots, t_{i_n j})$. The following development is based on van der Linden, Scrams, and Schnipke (1999). Assuming a normal prior for τ_j in the model in (4),

$$\tau_j \sim N(\mu_{0j}, \sigma_{0j}^2), \tag{11}$$

the posterior distribution of τ_j given $(t_{i_1 j}, \ldots, t_{i_n j})$ is also normal, with mean

$$\mathrm{E}(\tau_j \mid t_{i_1 j}, \ldots, t_{i_n j}) = \frac{\sigma^2 \mu_{0j} + \sigma_{0j}^2 \sum_{p=1}^{n} [\ln(t_{i_p j} - \mu - \delta_{i_p})]}{\sigma^2 + n\sigma_{0j}^2} \tag{12}$$

and variance

$$\mathrm{Var}(\tau_j \mid t_{i_1 j}, \ldots, t_{i_n j}) = \frac{\sigma_{0j}^2 \sigma^2}{\sigma^2 + n\sigma_{0j}^2}. \tag{13}$$

For a population of exchangeable examinees, the parameters in (11) can be chosen to be equal to the mean and variance of the empirical distribution of τ. From (5), it follows that for all j

$$\mu_{0j} = 0, \tag{14}$$

where

$$\sigma_{0j}^2 = \sigma_\tau^2. \tag{15}$$

Consequently, (12) and (13) specialize to

$$\mathrm{E}(\tau_j \mid t_{i_1 j}, \ldots, t_{i_n j}) = \frac{\sigma_\tau^2 \sum_{p=1}^{n} [\ln(t_{i_p j} - \mu - \delta_{i_p})]}{\sigma^2 + n\sigma_\tau^2} \tag{16}$$

and

$$\mathrm{Var}(\tau_j \mid t_{i_1 j}, \ldots, t_{i_n j}) = \frac{\sigma_\tau^2 \sigma^2}{\sigma^2 + n\sigma_\tau^2}. \tag{17}$$

Cross-Validation Residuals

All residuals for item i in this paper are calculated with the values of the examinee parameters θ or τ estimated only from the other items in the test. Such residuals are known as cross-validated residuals in the literature on regression models for binary data (e.g., Johnson & Albert, 1999, chap. 3). The correction prevents a possible bias in their size due the fact that the suspicious data are included in the estimation procedure. The same practice is recommended in the literature on residuals in linear models, where such the residuals are known as deleted residuals (or studentized deleted residuals if they are also studentized; see Neter, Wasserman, & Kutner, 1985, sec. 11.5). The vector of responses for examinee j omitting the response to the k-th item in the test is denoted as $\mathbf{U}_{(i_k)j}$, the vector of response times as $\mathbf{T}_{(i_k)j}$. Likewise, the estimators of θ and τ for examinee j without the k-th item are denoted as $\widehat{\theta}_{(i_k)j}$ and $\widehat{\tau}_{(i_k)j}$.

Predicted Distribution of Response Times

Let $\widetilde{T}_{i_k j}$ denote the predictor of the response time distribution for examinee j on the k-th item in the test based on the MLE of $\tau_{(i_k)j}$. From (8) it follows that

$$\ln \widetilde{T}_{i_k j} \equiv \mu + \delta_{i_k} + \widehat{\tau}_{(i_k)j} + \varepsilon_{i_k j}. \tag{18}$$

From (9) and (10), for a known value of σ^2,

$$\widehat{\tau}_{(i_k)j} \sim N(\tau_j, \sigma^2/(n-1)), \tag{19}$$

whereas, from (3), $\varepsilon_{i_k j} \sim N(0, \sigma^2)$. Because

$$\mathrm{Cov}(\widehat{\tau}_{(i_k)j}, \varepsilon_{i_k j}) = 0, \tag{20}$$

it thus holds for $\ln \widetilde{T}_{i_k j}$ that

$$\ln \widetilde{T}_{i_k j} \sim N(\mu + \delta_{i_k} + \tau_j, n\sigma^2/(n-1)). \tag{21}$$

A Bayesian prediction of the response time distribution for examinee j on the k-th item is provided by the posterior predictive density of $\ln \widetilde{T}_{i_k j}$ given $\mathbf{t}_{(i_k)j}$. For the model in (4), with the normal prior for τ_j with the parameters in (14) and (15), the posterior predictive density is normal with mean

$$\mathrm{E}(\ln \widetilde{T}_{i_k j} \mid \mathbf{t}_{(i_k)j}) = \mu + \delta_{i_k} + \frac{\sum_{p=1, p\neq k}^{n} (\ln t_{i_p j} - \mu - \delta_{i_p})}{\sigma^2/\sigma_\tau^2 + n - 1} \tag{22}$$

and variance

$$\mathrm{Var}(\ln \widetilde{T}_{i_k j} \mid \mathbf{t}_{(i_k)j}) = \frac{\sigma^2 + n\sigma_\tau^2}{1 + (n-1)\sigma^2/\sigma_\tau^2}. \tag{23}$$

Classical and Bayesian Checks on Response Times

The cross-validated residual of the response time of examinee j on item k in the test is the difference between the predicted and the actual response time. On the logarithmic scale in (2), the residual is defined as

$$E_{i_k j} \equiv \ln \frac{\widetilde{T}_{i_k j}}{t_{i_k j}}. \tag{24}$$

A classical check on the realization $E_{i_k j} = e_{i_k j}$ can be based on the predicted distribution of $\ln \widetilde{T}_{i_k j}$ in (21). For this choice the residual is distributed as

$$E_{i_k j} \sim N(\mu + \delta_{i_k} + \tau_j - \ln t_{i_k j}, n\sigma^2/(n-1)). \tag{25}$$

Thus, an approximate Gaussian test on $e_{i_k j}$ is to compare its actual value with critical values under the normal distribution in (25) with the $\widehat{\tau}_{(i_k)j}$ substituted for τ_j. From (19) it follows that $\widehat{\tau}_{(i_k)j}$ is unbiased and converges in distribution to τ_j. Because its variance decreases in n, satisfactory approximation is expected for application with adaptive tests which typically have a length of 30 items or more.

A Bayesian posterior predictive check on $e_{i_k j}$ is based on the distribution of $\ln \widetilde{T}_{i_k j}$ with the parameters in (22) and (23) (for a general treatment of posterior predictive checks, see Gelman, Carlin, Stern, & Rubin, 1995). The distribution of $E_{i_k j}$ given $\mathbf{t}_{(i_k)j}$ is normal with mean

$$\mathrm{E}(E_{i_k j} \mid \mathbf{t}_{(i_k)j}) = \mu + \delta_{i_k} - \ln t_{i_k j} + \frac{\sum_{p=1, p\neq k}^{n} (\ln t_{i_p j} - \mu - \delta_{i_p})}{\sigma^2/\sigma_\tau^2 + n - 1} \tag{26}$$

and variance

$$\mathrm{Var}(E_{i_k j} \mid \mathbf{t}_{(i_k)j}) = \frac{\sigma^2 + n\sigma_\tau^2}{1 + (n-1)\sigma^2/\sigma_\tau^2}. \tag{27}$$

Unlike (25), the uncertainty due to the fact that τ_j is estimated from the actual response times on the other items is absorbed in the parameters of the posterior predictive distribution of $E_{i_j k}$. This distribution gives us exact critical values. In a Bayesian residual check, the value of $e_{i_j k}$ is compared with critical values under the normal distribution with the mean and variance in (26) and (27).

Classical and Bayesian Checks on Responses

Analogously to (24), the cross-validated residual of the response of examinee j on item k in the test is defined as the difference between the predicted response $\widetilde{U}_{i_k j}$ and actual response $u_{i_k j}$:

$$E_{i_k j} \equiv \widetilde{U}_{i_k j} - u_{i_k j}. \tag{28}$$

The predicted response $\widetilde{U}_{i_k j}$ is a Bernoulli variable with probability of success $\pi_{i_k j}$ which is a function of the examinee parameter θ and the parameters of item i_k. It follows that residual $E_{i_k j}$ has probability function

$$p(e_{i_k j}) = \begin{cases} \pi_{i_k j}^{e_{i_k j}} (1 - \pi_{i_k j})^{1 - e_{i_k j}} & \text{if} \quad u_{i_k j} = 0 \\ \pi_{i_k j}^{1 - |e_{i_k j}|} (1 - \pi_{i_k j})^{|e_{i_k j}|} & \text{if} \quad u_{i_k j} = 1. \end{cases} \tag{29}$$

A classical check on a realization $e_{i_k j}$ would be based on the response model in (1). That is, the Bernoulli parameter would be equated to

$$\pi_{i_k j} = \Pr\{U_{i_k j} = 1\} \equiv p_{i_k}(\theta_j), \tag{30}$$

with $\widehat{\theta}_{(i_k) j}$ substituted for θ_j. In the empirical examples below, $\widehat{\theta}$ is chosen to be the EAP estimator in (7). This estimator is known to be consistent but has a small-sample bias toward the location of the prior distribution of θ . However, for the typical test length in large-scale adaptive testing programs, the approximation is believed to be satisfactory.

A Bayesian check can be based on the posterior predictive distribution of $\widetilde{U}_{i_k j}$ given $\mathbf{u}_{(i_k) j}$, which has probability function

$$p(\widetilde{u}_{i_k j} \mid \mathbf{u}_{(i_k) j}) = \int p(\widetilde{u}_{i_k j} \mid \theta) p(\theta \mid \mathbf{u}_{(i_k) j}) d\theta, \tag{31}$$

where $p(\widetilde{u}_{i_k j} \mid \theta)$ follows from the response model in (1) and posterior density $p(\theta \mid \mathbf{u}_{(i_k) j})$ is defined in (6). The residual based on this prediction is distributed according to (29) with

$$\pi_{i_k j} \equiv p(1 \mid \mathbf{u}_{(i_k) j}). \tag{32}$$

Observe that this distribution does not involve any point estimate of θ. The uncertainty on examinee j's value of θ is absorbed in the posterior predictive distribution. The probabilities calculated from (29) with (32) are thus exact.

Any residual with probability close to zero is unlikely under the model (and the prior distribution of θ) and therefore suspicious.

Applications

The power of these checks on response times in adaptive testing is investigated for aberrant behavior caused by two of the potential problems inherent in adaptive testing that have been discussed earlier:

1. *Preknowledge of some of the items in the pool.* If examinees share their knowledge of the items in the pool and an examinee knows some of the items prior to the test, a likely result for these items is the combination of unexpectedly correct responses with unexpectedly short response times. This result can happen for items at any position in the test
2. *Differential speededness of the test.* If items that require much time happen to be overrepresented in the tests for some of the examinees, a likely result is the combination of unexpectedly incorrect responses and unexpectedly short response times for the items at the end of the test.

Both cases involve one-sided checks on responses and response times. Because both variables are independent if the models in (1) and (4) fit the items for the population of examinees, the probabilities of Type I errors of these checks are also independent.

Method

Item Pool

The item pool was a 186-item pool for the CAT version of the Arithmetic Reasoning Test in the Armed Services Vocational Aptitude Battery (ASVAB). The test is described in Segall, Moreno, and Hetter (1997). The items fitted the response model in (1). To estimate the item and structural parameters in the model in (4), the response times for a sample of 38,357 examinees were available. The examinees in this sample had estimates of the values for the slowness parameter, τ, with a mean equal to zero (see Equation 5) and a standard deviation equal to .375. A study of the goodness of fit of the model to these data yielded a excellent results. Figure 1 shows a typical Q-Q plots of the observed and expected response times under a normal, lognormal, gamma, and Weibull distribution for the ASVAB items. For a full report on these and other results, see Schnipke and Scrams (1997) and van der Linden, Scrams and Schnipke (1999).

CAT Algorithm

Adaptive tests were simulated for examinees with $\theta = -2.0, -1.5, \ldots, 2.0$. The response times were simulated through random draws from the lognormal distribution in (4) with $\tau = -.6, -.3, \ldots, .6$. The series of values for τ was chosen in view of the above empirical standard deviation of $\widehat{\tau}$ for the ASVAB population of examinees. For each combination of values for (θ, τ) the number of simulated examinees was equal to 500.

Two different item-selection criteria were used. For the conditions with the classical checks, the items were selected applying the maximum-information criterion at the updated estimate of θ (van der Linden & Pashley, 2000, sec. 2.4.1). The estimator of θ was the EAP estimator in (7) with an uninformative prior over $[-4, 4]$. For the conditions with the Bayesian check, the items were selected according to the Bayesian criterion of minimum expected posterior variance (van der Linden & Pashley, 2000, sec. 3.3). In all conditions, the initial estimate of $\hat{\theta}$ was set equal to zero.

In the classical conditions, τ was estimated use the MLE in (18). No estimates of τ were needed for the Bayesian checks with the parameters in (22) and (23) because these are based on the normal posterior predictive density for the response times only.

A fixed-length stopping rule was applied. To assess the effects of test length, the length was set equal to $n = 21$ and 31.

Simulating Aberrant Responses and Response Times

The condition of item preknowledge was simulated for every fifth item in the CAT, that is, for $k = 1, 6, 11, \ldots$. It was assumed that when an item was compromised, the correct response was known. Thus, for each compromised item, the response was selected to be correct with

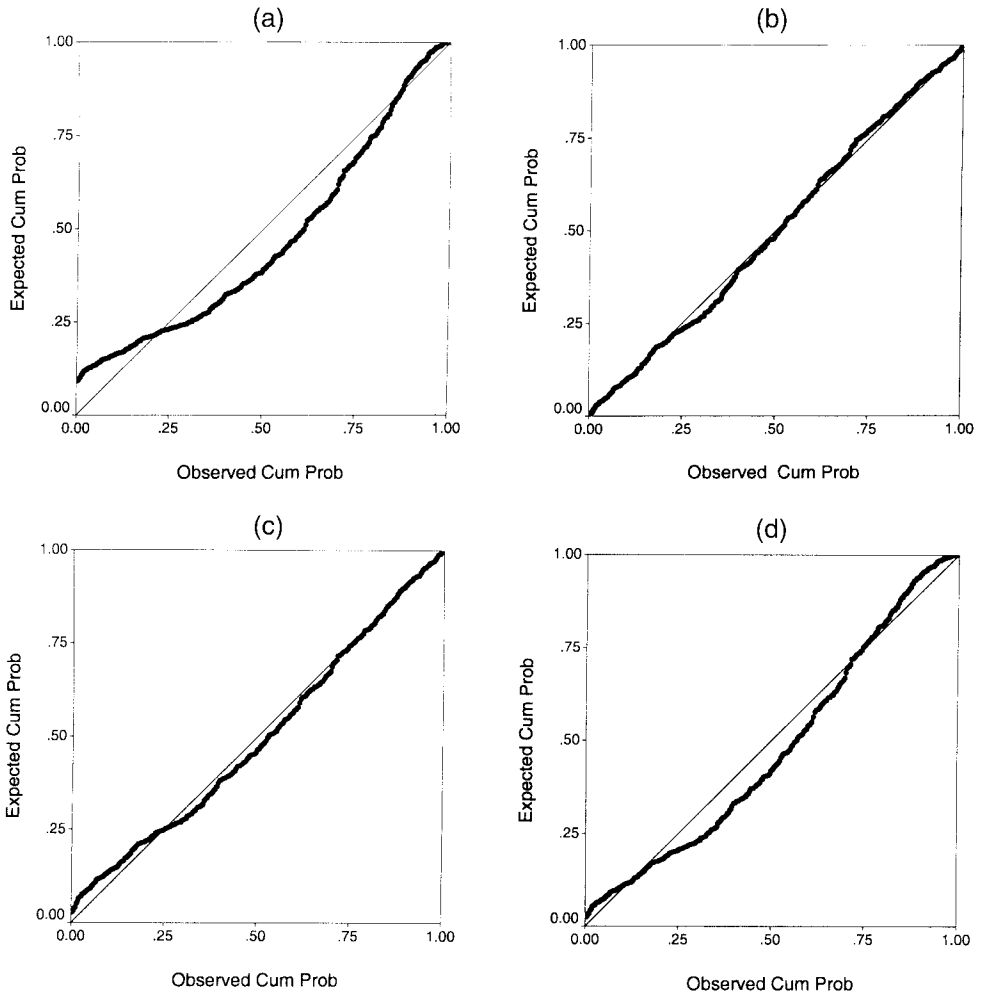

FIGURE 1.
Typical Q-Q plots of the observed and expected response times under (a) normal, (b) lognormal, (c) gamma, and (d) Weibull distribution for the ASVAB items.

probability 1. The logresponse times for the compromised items were sampled from

$$\ln T_{ij} \sim N(\mu + \delta_i + \tau_j + L, \sigma^2), \tag{33}$$

where L was a shift introduced by the experimenter. Both the effects of $L = -.375$ and $-.750$ (one and two times the empirical standard deviation of $\widehat{\tau}$ in the sample of ASVAB examinees, respectively) were examined.

The condition of differential speededness was simulated to have impact on the last five items in the test. For these examinees, the responses were simulated at $\theta' = \theta - r$ whereas their response times were drawn from (33). The combined effects of $r = 1$ and $L = -.375$ as well as $r = 2$ and $L = -.750$ were examined.

Calculation of Checks

One-sided checks with a level of significance $\alpha = .05$ were executed. The posterior predictive density in the Bayesian checks on the responses in (31) was calculated using numerical quadrature with 30 points.

Evaluation Measures

For each check the detection and false-alarm rates were estimated. Estimates of the detection rates were calculated as the proportion of simulees with responses (and response times) on the item flagged correctly as aberrant. Estimates of the false-alarm rates were calculated as the proportion of simulees with responses (and response times) on the item flagged erroneously as aberrant.

Results

The results for the classical checks on the responses as well as the response times for the two cases of aberrances are presented in Figure 2 ($n = 21$) and 3 ($n = 31$). The detection rates for the checks on the responses associated with the normal case and the two levels of preknowledge (Panel a) and speededness (Panel c) did show no power whatsoever. Note that the detection rates for the normal case are actually false-alarm rates. Also, the length of the test did not have any impact on these results.

The classical checks on the response times had results according to our expectations, both in the case of preknowledge (Panel b) and speededness (Panel d). The normal case showed a false-alarm close to .05. The two cases of aberrance had detection rates close to .15 for responses with the lower level of aberrance and .30 for responses with a higher level. The checks thus nicely discriminated between the items on which the examinees showed aberrant and normal response times. Again, the length of test had hardly any impact on the results, except for a slight increase in detection rate at the higher level of speededness.

The results for the Bayesian checks are in Figure 4 ($n = 21$) and 5 ($n = 31$). Though the general patterns in the detection rates resembles those in the two previous figures, a few idiosyncrasies, which are all Bayesian by nature, should be noted. First, the checks on the responses had some power to detect preknowledge of the first item in the test (Panel a). This result is due to the impact of the prior for θ on the predictive distribution for the responses. Because the prior was chosen to be uniform over $[-4, 4]$, its only effect was to bound the estimator of θ. Second, the detection rates for the aberrant response times almost doubled relative to the classical checks, but at the cost of a considerable increase in the false-alarm rate. This phenomenon is also typical of Bayesian analysis and due to the use of the informative prior for τ in (11).

The results in all figures were obtained pooling over the values of θ and τ simulated. However, analyses were also done for the individual combinations of parameter values. Because these analyses revealed no systematic differences between detection rates and the number of combinations were large, the separate results are not presented.

Discussion

The results from these empirical examples generally confirmed our expectations. The only exception was that the results for the checks on the responses were more discouraging than anticipated. In fact, using these checks at the level of individual responses can not be recommended at all.

Other types of aberrance than those studied here are possible. Examples are a warming-up effect for the examinee, misunderstanding of the instructions, or fatigue built up toward the end of the test. Each of these examples has another type of response pattern. Though not studied, it seems not too risky to assume that the results for these cases would have been comparable.

An important distinction is between types of aberrant examinee behavior that are fraudulent or due to a bad design of the test. It seems prudent to consider checks on responses and response times only as support to an hypothesis of fraud by an individual examinee and not as decisive evidence. On the other hand, evidence of bad test design can be based on larger series of response times within a test and larger samples of examinees and is generally much stronger.

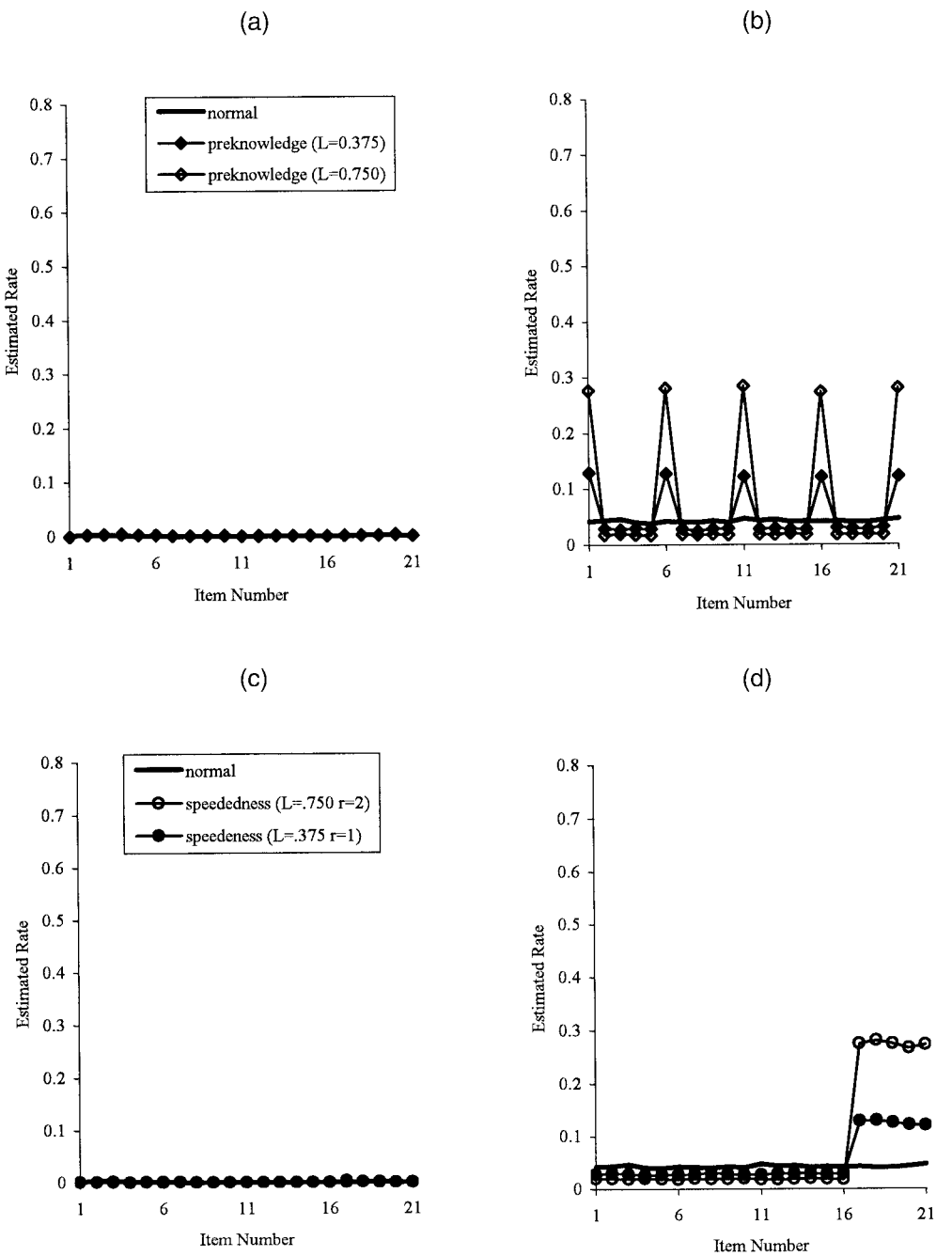

FIGURE 2.
Detection rates for the classical checks on responses and response times for the cases of preknowledge (Panels a and c) and speededness (Panels c and d) ($n = 21$).

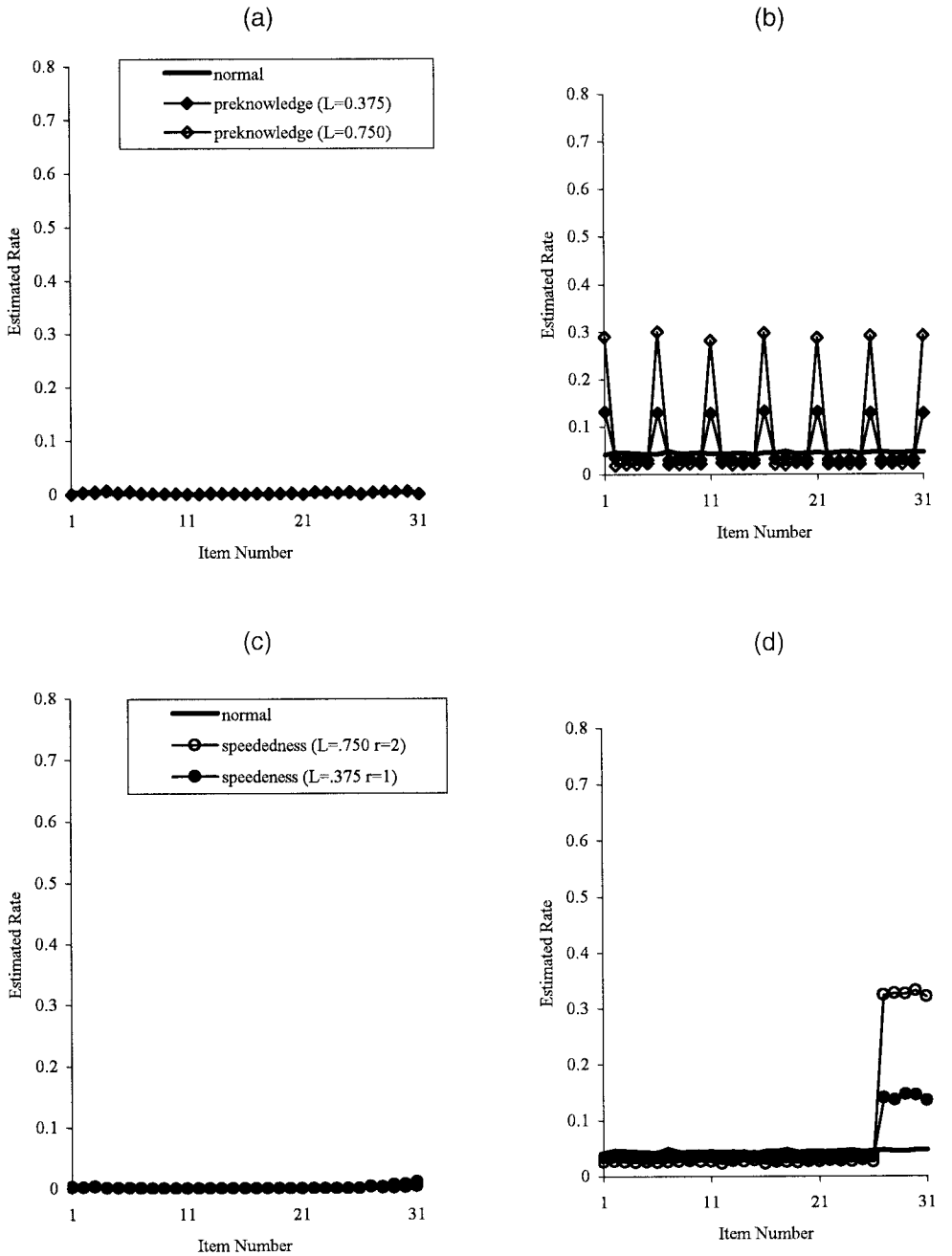

FIGURE 3.
Detection rates for the classical checks on responses and response times for the cases of preknowledge (Panels a and c) and speededness (Panels c and d) ($n = 31$).

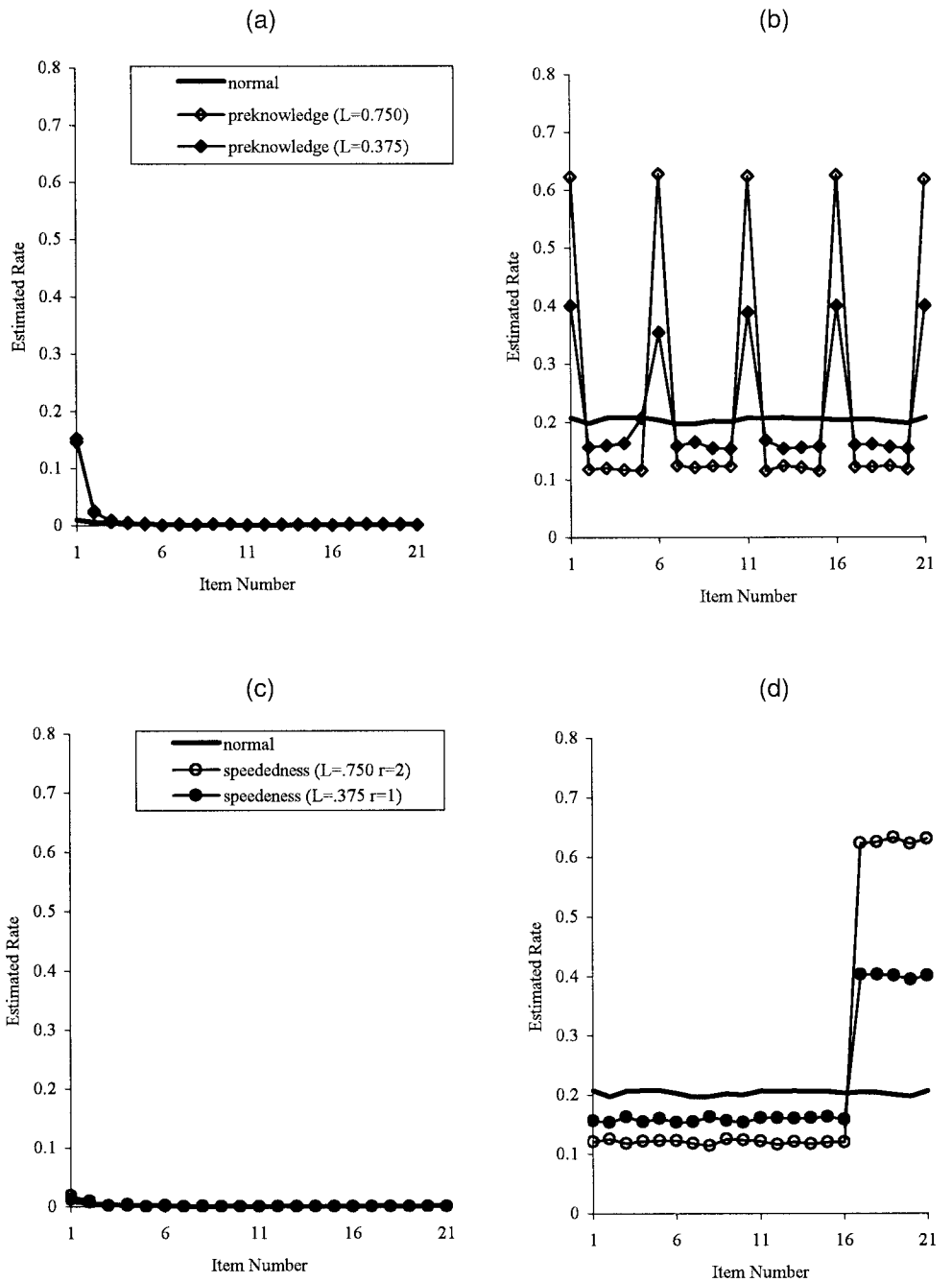

FIGURE 4.
Detection rates for the Bayesian checks on responses and response times for the cases of preknowledge (Panels a and c) and speededness (Panels c and d) ($n = 21$).

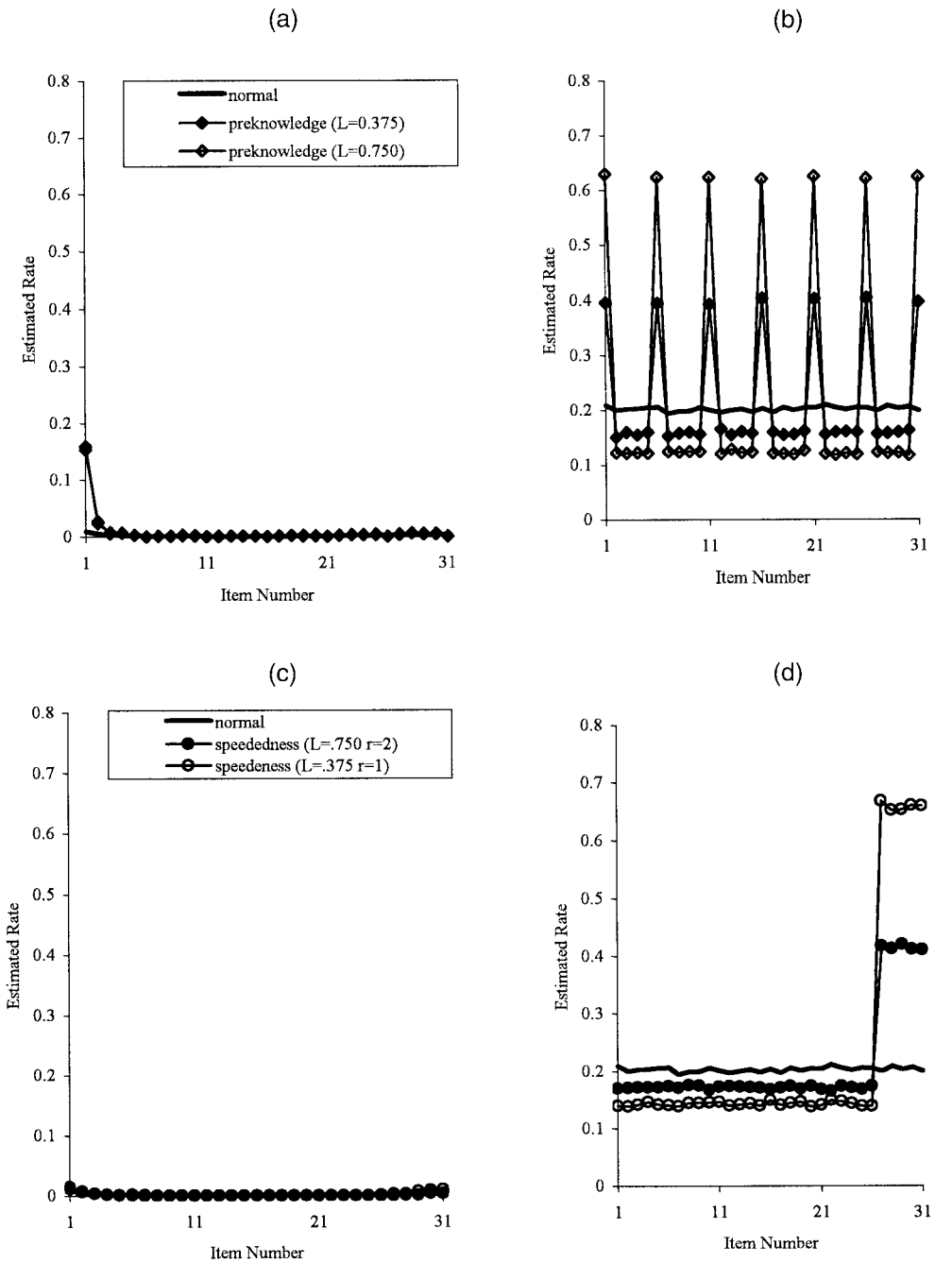

FIGURE 5.
Detection rates for the Bayesian checks on responses and response times for the cases of preknowledge (Panels a and c) and speededness (Panels c and d) ($n = 31$).

The distinction between these two types of aberrant behavior also provides a guideline for the choice between the types of checks on response times studied in this paper. If the goal is to detect a flaw in the design of the test, the high detection rates for the Bayesian checks with an informative prior are welcome. However, if fraudulent behavior among the examinees is surmised, control of the false-alarm rate should have a higher priority. In this case, the classical check or the Bayesian check with a less informative prior should be preferred.

References

Bradlow, E.T., Weiss, R. E., & Cho, M. (1998). Bayesian detection of outliers in computerized adaptive tests. *Journal of the American Statistical Association, 93*, 910–919.

Drasgow, F., Levine, M.V., Williams, E.A. (1985). Appropriateness measurement with polytomous item reponse models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67–86.

Gelman, A., Carlin, J.B, Stern, H., & Rubin, D.B. (1995). *Bayesian data analysis*. London, U.K.: Chapman & Hall.

Johnson, V.E., & Albert, J.H. (1999). *Ordinal data modeling*. New York, NY: Springer-Verlag.

Levine, M.V., & Rubin, D.B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269–290.

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

Mislevy, R.J., & Chang, H. (2000). Does adaptive testing violate local independence? *Psychometrika, 65*, 149–156.

Mislevy, R.J., & Wu, P.-K. (1996). *Missing responses and Bayesian IRT estimation: Omits, choice, time limits, and adaptive testing* (Research Rep. RR-96-30-ONR). Princeton, NJ: Educational Testing Service.

Molenaar, I.W., & Hoijtink, H. (1990). The many null distributions of person-fit statistics. *Psychometrika, 55*, 75–106.

Neter, J., Wasserman, W., & Kutner, M.H. (1985). *Applied linear statistical models: Regression, analysis of variance, and experimental designs*. Homewood, IL: Richard D. Irwin.

Segall, D.O., Moreno, K.E., & Hetter, D.H. (1997). In W.A. Sands, B.K. Waters, & J.R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 117–130). Washington, DC: American Psychological Association.

Schnipke, D.L., & Scrams, D.J. (1997). *Representing response time information in item banks* (LSAC Computerized Testing Rep. No. 97–09). Newtown, PA: Law School Admission Council.

Thissen, D. (1983). Timed testing: An approach using item resonse theory. In D.J. Weiss (Ed.), New horizons in testing: Latent trait test theory and computerized adaptive testing (pp. 179–203). New York, NY: Academic Press.

Trabin, T.E., & Weiss, D.J. (1983). The person response curve: Fit of individuals to item response theory models. In D.J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing*. New York, NY: Academic Press.

van der Linden, W.J. (2002). *A model for speed and accuracy on tests*. Unpublished manuscipt.

van der Linden, W.J., & Pashley, P.J. (2000). Item selection and ability estimation in adaptive testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1–25). Norwell, MA: Kluwer Academic Publishers.

van der Linden, W.J., Scrams, D.J., & Schnipke, D.L. (1999). Using response-time constraints to control for speededness in computerized adaptive testing. *Applied Psychological Measurement, 23*, 195–210.

van Krimpen-Stoop, E.M.L.A., & Meijer, R.R. (1999). Simulating the null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement, 23*, 327–345.

van Krimpen-Stoop, E.M.L.A., & Meijer, R.R. (2000). Detecting person misfit in adaptive testing using statistical process control techniques. In W.J. van der Linden & C.A.W. Glas. (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 221–219). Norwell, MA: Kluwer Academic Publishers.

Manuscript received 23 JAN 2001
Final version received 27 NOV 2001